Acquisition de liens sémantiques à partir d'éléments de mise en forme des textes

Jean-Philippe Fauconnier

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *27 janvier 2016* **par :**

Jean-Philippe Fauconnier

Acquisition de liens sémantiques à partir d'éléments de mise en forme des textes : exploitation des structures énumératives

---

JURY

Nathalie Aussenac-Gilles, Directrice de Recherche, CRNS/IRIT, Directrice de Thèse
Mouna Kamel, Maître de Conférences, Université de Perpignan, Directrice de Thèse
Thierry Poibeau, Directeur de Recherche, CNRS/LaTTiCe, Rapporteur
Pascale Sébillot, Professeur des Universités, INSA de Rennes/IRISA, Rapporteur
Béatrice Daille, Professeur des Universités, Université de Nantes/LINA, Présidente
Olivier Ferret, Ingénieur Chercheur, CEA LIST/LVIC, Examinateur
Núria Gala, Maître de Conférences, Université d'Aix-Marseille/LIF, Examinatrice

---

École doctorale et spécialité :
MITT : Domaine STIC : Intelligence Artificielle
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)

Ce document a été préparé avec LaTeX 2$_\varepsilon$ et une version modifiée de la classe classicthesis d'André Miede. La classe originale est accessible sur https://www.ctan.org/tex-archive/macros/latex/contrib/classicthesis/. Les polices utilisées sont Computer Modern Roman (cmr) et Latin Modern Sans Serif (lmss). Les différents arbres et graphes ont été réalisés avec les classes tikz et rst. L'éditeur de texte utilisé est l'éditeur libre Vi IMproved (Vim). Le code source de ce document ainsi que toutes les ressources associées sont accessibles sur demande et librement modifiables selon les termes de la licence Creative Commons BY-NC-SA 3.0.

Résumé

Ces dernières années de nombreux progrès ont été faits dans le domaine de l'extraction de relations à partir de textes, facilitant ainsi la construction de ressources lexicales ou sémantiques. Cependant, les méthodes proposées (apprentissage supervisé, méthodes à noyaux, apprentissage distant, etc.) n'exploitent pas tout le potentiel des textes : elles ont généralement été appliquées à un niveau phrastique, sans tenir compte des éléments de mise en forme.

Dans ce contexte, l'objectif de cette thèse est d'adapter ces méthodes à l'extraction de relations exprimées au-delà des frontières de la phrase. Pour cela, nous nous appuyons sur la sémantique véhiculée par les indices typographiques (puces, emphases, etc.) et dispositionnels (indentations visuelles, retours à la ligne, etc.), qui complètent des formulations strictement discursives. En particulier, nous étudions les structures énumératives verticales qui, bien qu'affichant des discontinuités entre leurs différents composants, présentent un tout sur le plan sémantique. Ces structures textuelles sont souvent révélatrices de relations hiérarchiques.

Notre travail est divisé en deux parties. (i) La première partie décrit un modèle pour représenter la structure hiérarchique des documents. Ce modèle se positionne dans la suite des modèles théoriques proposés pour rendre compte de l'architecture textuelle : une abstraction de la mise en forme et une connexion forte avec la structure rhétorique sont faites. Toutefois, notre modèle se démarque par une perspective d'analyse automatique des textes. Nous en proposons une implémentation efficace sous la forme d'une méthode ascendante et nous l'évaluons sur un corpus de documents PDF.

(ii) La seconde partie porte sur l'intégration de ce modèle dans le processus d'extraction de relations. Plus particulièrement, nous nous sommes focalisés sur les structures énumératives verticales. Un corpus a été annoté selon une typologie multi-dimensionnelle permettant de caractériser et de cibler les structures énumératives verticales porteuses de relations utiles à la création de ressources. Les observations faites en corpus ont conduit à procéder en deux étapes par apprentissage supervisé pour analyser ces structures : qualifier la relation puis en extraire les arguments. L'évaluation de cette méthode montre que l'exploitation de la mise en forme, combinée à un faisceau d'indices lexico-syntaxiques, améliore les résultats.

Abstract

The past decade witnessed significant advances in the field of relation extraction from text, facilitating the building of lexical or semantic resources. However, the methods proposed so far (supervised learning, kernel methods, distant supervision, etc.) don't fully exploit the texts : they are usually applied at the sentential level and they don't take into account the layout and the formatting of texts.

In such a context, this thesis aims at expanding those methods and makes them layout-aware for extracting relations expressed beyond sentence boundaries. For this purpose, we rely on the semantics conveyed by typographical (bullets, emphasis, etc.) and dispositional (visual indentations, carriage returns, etc.) features. Those features often substitute purely discursive formulations. In particular, the study reported here is dealing with the relations carried by the vertical enumerative structures. Although they display discontinuities between their various components, the enumerative structures can be dealt as a whole at the semantic level. They form textual structures prone to hierarchical relations.

This study was divided into two parts. (i) The first part describes a model representing the hierarchical structure of documents. This model is falling within the theoretical framework representing the textual architecture : an abstraction of the layout and the formatting, as well as a strong connection with the rhetorical structure are achieved. However, our model focuses primarily on the efficiency of the analysis process rather than on the expressiveness of the representation. A bottom-up method intended for building automatically this model is presented and evaluated on a corpus of PDF documents.

(ii) The second part aims at integrating this model into the process of relation extraction. In particular, we focused on vertical enumerative structures. A multidimensional typology intended for characterizing those structures was established and used into an annotation task. Thanks to corpus-based observations, we proposed a two-step method, by supervised learning, for qualifying the nature of the relation and identifying its arguments. The evaluation of our method showed that exploiting the formatting and the layout of documents, in combination with standard lexico-syntactic features, improves those two tasks.

Remerciements

Car une thèse de doctorat est également une aventure humaine, j'aimerais remercier un grand nombre de personnes sans lesquelles ce travail, et le manuscrit résultant, n'auraient pas pu voir le jour. Je leur dois beaucoup.

En premier lieu, j'adresse mes vifs et sincères remerciements à mes Directrices de Recherche. Mouna Kamel, pour la confiance qu'elle m'a accordée dès le début de ce travail en 2012, ainsi que pour m'avoir encouragé à donner le meilleur de moi-même au cours des années. Nathalie Aussenac-Gilles, pour son appui scientifique, sa vision claire du domaine et sa gentillesse. J'imagine la difficulté liée à l'exercice d'encadrer un doctorant. Pour tout cela, je vous remercie encore une fois toutes les deux.

Ensuite, j'aimerais remercier mes Rapporteurs Madame Pascale Sébillot et Monsieur Thierry Poibeau pour avoir accepté d'évaluer mon travail. La pertinence de vos remarques et la clarté de vos propos ont été très inspirantes et m'ont permis d'améliorer ce travail. Je tiens également à remercier mes Examinateurs Madame Béatrice Daille, Monsieur Olivier Ferret et Madame Núria Gala, pour avoir accepté de participer à mon jury. La qualité et la teneur des échanges lors de la soutenance ont en fait un moment scientifiquement et humainement vivifiant.

Ce travail n'aurait pas été possible sans Cécile Fabre et Ludovic Tanguy. Un grand merci pour tout ce que vous m'avez appris ainsi que votre support dans le procédé de sélection des candidats. De même, je remercie Pieter Vankeerberghen, Joseph Roumier et Fabrice Estiévenart pour avoir aussi appuyé ma candidature. Merci à vous. Mes remerciements s'adressent également à l'ensemble de l'équipe MELODI au sein de laquelle j'ai trouvé un milieu favorable au questionnement scientifique. J'ai beaucoup appris parmi vous. Un grand merci pour cela.

J'aimerais remercier Bernard Rothenburger, pour nos nombreuses collaborations scientifiques. Mustapha Mojahid, pour ses idées éclairantes en Structure de Document. Laure Vieu, pour nos échanges en Sémantique Lexicale. Véronique Moriceau, pour son aide sur Kitten et Citron. Mai Ho-Dac, à la fois pour m'avoir enseigné avec autant de passion le TAL lorsque j'étais étudiant, mais également pour tous ses conseils lors de ma thèse. Thomas François, notamment pour m'avoir invité à venir présenter mes travaux. Tim Van de Cruys, pour nos conversations autour de la bière belge et des réseaux de neurones. Assaf Urieli, pour m'avoir permis de travailler sur Talismane. J'ai littéralement acquis les bases de l'apprentissage supervisé à tes côtés. Je tiens également à remercier Marco Serranos avec qui j'ai eu l'occasion d'enseigner, et qui a soutenu ma candidature

ATER. Martine Labruyère, pour sa patience et sa gentillesse. Nathalie Hernandez et Cassia Trojhan, mes nouvelles collègues de bureau, pour leurs encouragements dans les derniers mois. Un grand merci à tous.

Une pensée pour mes comparses doctorants de l'IRIT : Camille Pradel, Julien Corman, Antoine Venant et Morgane de Coninck, Antoine Bride, Juliette Conrath, Nicolas Seydoux, Jihen Karoui, Rafik Abbes, Jeremy Perret, Fabien Amarger, Pierre Bisquert, François Gatto, Nadine Guiraud, Laurent Sorin, Anaïs Cadilhac. Une pensée également pour ceux rencontrés à l'UT2J : François Morlane-Hondère, Simon Leva, Caroline Atallah, Marianne Vergez-Couret, Cécile Viollain, Clémentine Adam pour ne citer qu'eux. En particulier, je remercie les membres de l'équipe JeTou 2015 : Maxime Warnier, Luce Lefevre, Laury Garnier, Olivier Nocaudie, Florian Savreux, Francesca Cortelazzo. Sans oublier mes acolytes belges et leur amitié malgré le temps et la distance : Mad Tihon, Gauthier Wilmet, Joachim Soudan, Renato Luna, Nicolas Vanstalle, Alex Kovalev, Ouliana Tolstova, Guillaume Uyttersprot, Nathan Gurnet, Damien Bouilliez, Jérôme Van Den Broeck, et les moins belges Anne Schwab et Héloïse Terrats. Un remerciement spécial pour Fanny Saintes et ses relectures assidues.

La Famille aussi est primordiale. Un mot d'abord pour ma belle famille qui m'a énormément encouragé durant les derniers mois. Je ne peux pas tous les citer ici, mais ils se reconnaîtront. Un grand merci à vous tous. Ensuite, Je remercie ma grand-mère Mamy, utilisatrice chevronnée de Linux, pour son écoute et ses encouragements. Une pensée également pour mon grand-père, parti un peu avant la fin. J'aime à penser qu'il aurait été fier d'avoir un petit-fils docteur. Un mot pour ma grand-mère Boma également. Ensuite, viennent mes parents : Merci pour votre soutien indéfectible ! Un signe aux frères et à la sœur : Maxime, Charlotte et Xavier, Bruno, Pierrick. Enfin, merci Valérie pour ta patience et ton attention.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'extraction de relations constitue un enjeu majeur pour l'acquisition de connaissances à partir de textes. L'objectif de cette tâche est de détecter et de caractériser des relations entre des entités textuelles (Murphy, 2003). Une fois extraites, ces relations peuvent être employées dans un processus de construction de ressources lexicales ou sémantiques, rendant possible une représentation des connaissances sous une forme exploitable par une machine (Pantel et Pennacchiotti, 2008).

De nombreuses approches ont été proposées pour extraire des relations à partir des textes. Trois grandes catégories peuvent être considérées : les approches symboliques, les approches statistiques et, enfin, les approches hybrides combinant les deux premières. L'exemple (0.a), extrait de Wikipédia[1], montre une phrase que ces types d'approches peuvent exploiter afin d'extraire une relation d'hyperonymie entre *ouvrage d'art hydraulique* et *écluse*.

(0.a)	Une écluse est un ouvrage d'art hydraulique implanté dans un canal ou un cours d'eau pour le rendre navigable et permettre aux bateaux de franchir des dénivellations.

Cependant, ces approches n'exploitent pas tout le potentiel des textes : elles ont généralement été appliquées à un niveau phrastique, sans tenir compte des éléments de mise en forme.

Dans ce travail, nous soutenons la thèse que des indices typographiques et dispositionnels peuvent être exploités pour découvrir de nouvelles relations, hors d'atteinte des approches classiques. Nous pensons que ces indices peuvent signaler, au même titre que des indices lexico-syntaxiques avec lesquels ils se conjuguent, des phénomènes sémantiques à l'échelle du texte, incluant notamment l'expression de relations sémantiques.

En particulier, notre travail s'intéresse aux structures énumératives (SE), qui constituent un terrain idéal pour l'étude des interactions entre mise en forme et richesse sémantique (Porhiel, 2007). Du point de vue de leur réalisation, les SE mettent en œuvre

[1] Exemple extrait de la page `http://fr.wikipedia.org/wiki/Écluse` (*dump* 2014-09-28).

différents mécanismes : elles passent de formes linéaires, dites *horizontales*, réalisées au travers de constructions syntaxiques (juxtaposition, coordination, etc.), à des formes matérialisées typographiquement et dispositionnellement, dites *verticales*, qui les rendent perceptibles à la surface des textes. Du point de vue sémantique, ces structures textuelles sont intéressantes, car elles sont propices aux relations sémantiques hiérarchiques, utiles à la création de ressources.

Dans ce travail, nous ciblons les SE verticales, dont le marquage visuel permet d'envisager leur identification dans les textes mais également le bornage de leurs composants internes. L'exemple (0.b), extrait de Wikipédia[2], montre une SE verticale porteuse d'une relation d'hyperonymie distribuée entre *navires de services* et, respectivement, *dragues*, *bateaux pilote*, *remorqueurs portuaires* et *bateaux de lamanage*.

(0.b)

Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets…) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants. Les matériaux extraits sont transportés par une marie-salope.
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.

S'appuyer sur des éléments de mise en forme pour dégager des structures textuelles amène néanmoins de nouvelles difficultés : la variabilité des indices visuels ainsi que la diversité des pratiques de formatage empêchent une prise en compte directe de ceux-ci (Pascual et Virbel, 1996). Pour permettre une approche générique et indépendante du format d'entrée, il est nécessaire d'avoir recours à une certaine forme d'abstraction.

Des difficultés similaires apparaissent dans l'analyse des SE. Déterminer si les relations portées par celles-ci sont utiles à la construction de ressources et, cas échéant, extraire les arguments implique de jongler avec des indices de nature et de couverture différentes. Une réflexion sur ceux-ci et une analyse linguistique en corpus sont préalables.

[2] Exemple extrait de la page `http://fr.wikipedia.org/wiki/Port` (*dump* 2014-09-28). Pour la clarté de l'exemple, les derniers items ont été enlevés. L'exemple complet est accessible en Annexe D.1.

Objectifs

Au regard de la problématique énoncée et des difficultés associées, nous présentons ci-dessous nos objectifs :

1. proposer un modèle qui permette l'abstraction de la structure des documents au travers des éléments de mise en forme qu'elle sollicite.

2. développer un système qui implémente ce modèle et qui permette une analyse automatique pour un document donné ;

3. proposer une analyse linguistique des relations sémantiques portées par les structures énumératives sur des exemples attestés en corpus afin de mettre au jour des indices informatifs et des liens entre ceux-ci ;

4. développer un système qui prenne en compte le modèle de structure de document et les indices linguistiques vus en corpus afin d'extraire les relations sémantiques à partir de structures énumératives verticales.

Cadre méthodologique

Afin de mener à bien les objectifs précités, nous inscrivons notre démarche dans une approche à l'échelle du document (Péry-Woodley et Scott, 2006). Le terme de document reflète ici un texte ancré dans un contexte de production donné et matériellement réalisé au travers d'un dispositif (au sens étendu : crayon, machine à écrire, outil de composition, etc.). Dans ce contexte, nous pensons que les indices typo-dispositionnels participent à la structure de cohérence, au même titre que les indices de cohésions référentielle, relationnelle et lexicale (Morris et Hirst, 1991). La mise au jour de cette structure de cohérence (la « texture » de Halliday (1977)) nous semble essentielle pour l'amélioration des systèmes d'extraction d'information, et en particulier le nôtre. Ce choix nous rapproche du travail de Couto *et al.* (2004), que nous étendons en rendant compte d'éléments de mise en forme. *A contrario*, nous nous démarquons des approches abordant les textes au travers de ce que Nazarenko (2005) appelle des « îlots textuels ».

La difficulté de faire émerger des indices discriminants lorsqu'aucun n'est pleinement suffisant ou nécessaire pour signaler un phénomène linguistique donné, nous a fait considérer la nécessité d'une approche par apprentissage statistique afin de prendre en compte des « faisceaux d'indices » (Ho-Dac *et al.*, 2009). Toutefois, nous pensons que l'utilisation d'apprentissage statistique pour ce type de problème ne peut être exécutée sans une supervision humaine forte. Le choix des indices initialement pris en compte nécessite une réflexion linguistique sur des exemples attestés en corpus, ainsi qu'une annotation manuelle et fine de ceux-ci. Ce choix rejoint la voie méthodologique proposée par Biber *et al.* (2007), et empruntée par Laignelet (2009).

Notre système se présente sous la forme d'une suite modulable d'outils prenant en entrée des documents et fournissant en sortie les relations extraites des structures énumératives verticales qu'ils présentent. La figure 0.1 schématise cette chaîne.

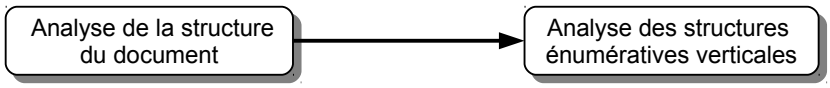

FIGURE 0.1 : Chaîne de traitement pour l'extraction de relations sémantiques présentes dans les structures énumératives verticales

Publications

Les réflexions et les résultats de cette thèse apparaissent dans les articles suivants :

FAUCONNIER, J.-P. et KAMEL, M. (2015). Discovering hypernymy relations using text layout. *In Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 249–258, Denver, Colorado. Association for Computational Linguistics.

FAUCONNIER, J.-P., KAMEL, M. et ROTHENBURGER, B. (2015). A Supervised Machine Learning Approach for Taxonomic Relation Recognition through Non-linear Enumerative Structures (papier court). *In ACM Symposium on Applied Computing (SAC 2015)*, Salamanque.

KAMEL, M., ROTHENBURGER, B. et FAUCONNIER, J.-P. (2014). Identification de relations sémantiques portées par les structures énumératives paradigmatiques. *Revue d'Intelligence Artificielle*, Ingénierie des Connaissances.

FAUCONNIER, J.-P., SORIN, L., KAMEL, M., MOJAHID, M. et AUSSENAC-GILLES, N. (2014). Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux. *In Actes de la 21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, pages 340–351.

FAUCONNIER, J.-P., KAMEL, M. et ROTHENBURGER, B. (2013a). Une typologie multidimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques. *In International Conference on Terminology and Artificial Intelligence (TIA 2013)*.

FAUCONNIER, J.-P., KAMEL, M., ROTHENBURGER, B. et AUSSENAC-GILLES, N. (2013b). Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles. *In Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, pages 132–145.

Plan du manuscrit

Dans la suite de ce manuscrit, nous développons les étapes qui ont conduit aux propositions de réponses des objectifs, selon le cadre méthodologique avancé :

Partie I : Contexte de l'étude

- **Chapitre 1** : ce chapitre est le chapeau du manuscrit. Nous revenons sur la tâche d'extraction de relations à partir de textes. La classification que nous proposons distingue les approches opérant sur du contenu textuel non-structuré et celles prenant en compte des éléments de mise en forme.

- **Chapitre 2** : ce chapitre est consacré à la structure de document. Nous décrivons trois modèles théoriques de structuration de document, issus du champ de la génération de textes, et nous présentons un travail original de comparaison entre ceux-ci. Sont ensuite décrits les approches empiriques liées à l'analyse automatique de la structure, ainsi que quelques formats où la structure intervient.

- **Chapitre 3** : ce chapitre introduit la structure énumérative, objet linguistique central de cette thèse. Nous présentons deux typologies sur lesquelles s'appuie notre travail. Ensuite, nous proposons un inventaire des approches exploitant, en le mentionnant explicitement ou non, les structures énumératives pour l'acquisition de connaissances.

Partie II : Modélisation et identification automatique de la structure de document

- **Chapitre 4** : ce chapitre propose un modèle de structure de document ouvrant la voie à la détection des structures énumératives marquées visuellement. Ce modèle s'inscrit dans la suite des modèles théoriques présentés dans le chapitre 2 : une abstraction de la mise en forme et une connexion forte avec l'aspect rhétorique sont proposées. Toutefois, notre modèle se démarque en se positionnant dans une perspective d'analyse, et non de générations de textes. Cela est notamment permis en ayant recours à une représentation en dépendances.

- **Chapitre 5** : ce chapitre présente une méthode qui implémente le modèle décrit dans le chapitre 4. Il s'agit d'une méthode ascendante partant de la forme visuelle du document pour aboutir à sa représentation en dépendances. Les algorithmes utilisés trouvent une correspondance avec ceux utilisés en parsing syntaxique et rhétorique. Nous donnons une évaluation sur un corpus de documents PDF.

Partie III : Extraction de relations dans les structures énumératives verticales

- **Chapitre 6** : ce chapitre propose une typologie pour caractériser et cibler les structures énumératives porteuses de relations sémantiques utiles à la construction de ressources. Cette typologie s'appuie sur les typologies présentées dans le chapitre 3, mais considère additionnellement une dimension sémantique. Sur la base de cette typologie, une campagne d'annotation a été menée. Un retour d'expérience et les résultats associés à la campagne sont donnés.

- **Chapitre 7** : ce chapitre présente une méthode pour l'extraction de relations à partir de structures énumératives verticales. La détection de ces dernières est permise grâce au modèle de structure de document et son implémentation présentés dans les chapitres 4 et 5. Les observations faites dans le corpus annoté du chapitre 6 ont conduit à procéder au travers de deux étapes : nous qualifions la relation sémantique portée, ensuite nous identifions les arguments de celle-ci. Chacune de ces deux étapes a été évaluée individuellement sur le corpus annoté, et l'ensemble du système a été évalué sur des données externes.

Première partie

Contexte de l'étude

Chapitre 1

Extraction de relations

Sommaire

Dans ce chapitre nous présentons un état de l'art sur la tâche d'extraction de relations sémantiques à partir de textes écrits en langage naturel. Notre attention se porte sur l'expression linguistique de ces relations et sur leur repérage en texte.

Le chapitre est structuré en deux parties. Dans un premier temps, nous délimitons la notion de relation sémantique et présentons le rôle que celle-ci joue dans quelques approches en sémantique lexicale. Également, nous décrivons la nature des entités textuelles qu'elle peut lier : les termes et les entités nommées. Dans un second temps, nous proposons une classification distinguant les approches d'extraction de relations qui utilisent du texte non structuré et celles qui combinent contenu textuel et éléments de mise en forme.

1.1 Positionnement théorique du problème

Lorsqu'il s'agit d'extraire des relations sémantiques à partir de textes, plusieurs questions doivent nécessairement être posées. Quel est l'intérêt de la tâche ? Quelle est la nature des relations recherchées ? Quelle est la nature des entités textuelles que celles-ci lient ?

1.1.1 Considérations générales

Une raison pour laquelle les relations sémantiques sont importantes est liée à leur rôle dans la représentation des connaissances (Feigenbaum et McCorduck, 1983). Les liens qu'elles tissent entre les entités constituent une ressource pour des applications de haut niveau. En outre, la représentation formelle qu'elles offrent ouvre la voie à l'inférence automatique de nouvelles connaissances. Dans les processus traitements textuels, leur extraction constitue généralement une des premières étapes (Green *et al.*, 2002).

Deux types de relations sémantiques sont généralement distingués (Sébillot, 2002) :

- les **relations paradigmatiques** qui regroupent les entités constitutives de classes ou directement liées sémantiquement (p. ex. *véhicule* et *voiture*) ;
- les **relations syntagmatiques** qui décrivent la combinatoire lexicale au travers des capacités d'association et de sélection des entités (p. ex. *garer* et *véhicule*).

Une autre distinction peut être faite entre *relation sémantique* et *relation lexicale* (Murphy, 2003). La première lie des *entités* (p. ex. mots ou segments de texte – clause, phrase, etc. –) dans des relations à caractère sémantique comme l'implication ou la contradiction. La seconde concerne les relations entre *unités lexicales*, au sens de Cruse (1986)[1], mais ces relations peuvent ne pas être sémantiques telles que les relations morphologiques (p. ex. la variation flexionnelle) ou les relations phonétiques (c'est-à-dire les rimes, les allitérations, etc.).

Dans ce document, nous centrons notre propos sur les relations paradigmatiques à caractère sémantique entre unités lexicales. Nous désignerons ce type de relations par le terme *relation sémantique*.

1.1.2 Sémantique lexicale et relations sémantiques

La sémantique lexicale s'intéresse aux questions liées à la place des relations sémantiques dans la construction du sens et à la distinction entre structures linguistiques et conceptuelles (Cruse, 2002). Par exemple, les relations prévalent-elles dans la définition d'une unité lexicale ? L'unité lexicale *chaud* est-elle définie lorsque son antonyme *froid* est connu ? Ou bien les propriétés attribuées à *chaud* permettent la dérivation d'une antonymie avec *froid* ? Les réponses à ces questions sont dépendantes des approches.

[1] Cruse oppose les *unités lexicales*, unités de sens (relativement) stable qui interagissent avec d'autres au travers de relations paradigmatiques et syntagmatiques, aux *lexèmes*, qui sont les éléments listés dans le lexique ou le « dictionnaire idéal » d'un langage Cruse (1986, p.49).

Nous reprenons ici la classification faite par Murphy (2003). Celle-ci organise les approches selon la place qu'elles accordent aux relations sémantiques. Trois catégories d'approches peuvent être distinguées :

- les **approches compositionnelles** où le sens d'une unité lexicale est représenté par des sous-composants et les relations sémantiques sont dérivées de ceux-ci ;
- les **approches associatives** où le sens d'une unité lexicale est représenté par des sous-composants et des relations sémantiques ;
- les **approches holistes et associatives** où le sens d'une unité lexicale est uniquement fonction des relations qui la lient à d'autres unités.

Nous les détaillons ci-dessous en donnant quelques exemples d'approches.

Approches compositionnelles Popularisées par la linguistique générative, les approches compositionnelles visent à représenter le sens d'une unité lexicale sous la forme d'une série de primitives ou de marqueurs sémantiques (pour *semantic marker*) (Katz et Fodor, 1963)[2]. Dans ce contexte, le sens d'une unité lexicale est défini par ses sous-composants.

Afin d'éviter le problème de la circularité des définitions, les approches compositionnelles proposent généralement un métalangage pour définir ces primitives (Katz et Fodor, 1963; Winograd, 1978; Jackendoff, 1990; Pustejovsky, 1995) ou bien utilisent un vocabulaire limité (Wierzbicka, 1972). Les relations sémantiques peuvent alors être dérivées à partir des définitions des unités lexicales. Katz et Fodor (1963) déclarent[3] :

> The semantic markers and distinguishers are the means by which we can decompose the meaning of one (s)ense of a lexical item into its atomic concepts, and thus exhibit the semantic structure IN a dictionary entry and the semantic relations BETWEEN dictionary entries. That is, the semantic relations among the various senses of a lexical item and among the various senses of different lexical items are represented by formal relations between markers and distinguishers.

Ce point de vue implique que les relations d'hyperonymie, de synonymie, etc. peuvent être définies en termes de similarité et de différences entre marqueurs sémantiques.

Jackendoff (1990) et Pustejovsky (1991; 1995) proposent des approches compositionnelles plus complexes, mais maintiennent toutefois une distinction claire entre le niveau des connaissances linguistiques et le niveau conceptuel. Pustejovsky (1991) le montre dans son lexique génératif :

> (...) the meanings of words should somehow reflect the deeper, conceptual structures in the system and the domain it operates in. This is tantamount to stating that the semantics of natural language should be the image of nonlinguistic conceptual organizing principles (whatever their structure).

[2] Ceci n'est pas surprenant, car Katz et Fodor ont travaillé avec Chomsky au MIT.
[3] Dans ce travail, les marques d'emphase présentes dans les citations sont celles des auteurs cités.

Approches associatives Pour représenter le sens d'une unité, les approches associatives utilisent à la fois une définition intralexicale et une représentation explicite des relations sémantiques. La définition intralexicale est justifiée soit pour son rôle de support additionnel de sens (Cruse, 1986), soit parce que les relations sémantiques représentées peuvent être arbitraires (Mel'čuk, 1988). Nous présentons ici trois perspectives : la sémantique des cadres, la théorie Sens-Texte et la position de Lyons et Cruse.

Sémantique des cadres Inspiré par l'intelligence artificielle (Minsky, 1975) et les travaux sur le champ lexical (Lehrer, 1974), Fillmore (1982) propose une approche du lexique reposant sur la notion de cadre cognitif. Les cadres sont des modèles où les signes linguistiques réfèrent à des catégories cognitives. Ces catégories sont organisées entre elles et participent à une structure conceptuelle plus large. Dans ce contexte, le sens donné à une unité l'est uniquement par son lien à un cadre, et les relations sémantiques ne sont pas considérées comme des objets linguistiques[4]. Notons que ce principe sera par la suite appliqué dans la création de la ressource FrameNet (Baker *et al.*, 1998).

Théorie Sens-Texte Dans le lexique de la théorie Sens-Texte (Mel'čuk, 1988), les entrées lexicales sont décrites par trois zones : la *zone syntaxique* qui contient les cadres de sous-catégorisation, la *zone sémantique* qui contient la définition compositionnelle de l'entrée et la *zone de co-occurrence lexicale* qui inclut toutes entrées liées par des relations syntagmatiques et paradigmatiques. Ces relations sont exprimées au travers de fonctions lexicales. Celles-ci prennent en argument une unité lexicale et en retournent une autre[5], avec néanmoins une certaine forme d'arbitraire[6] qui nécessite un recours à la définition compositionnelle. Notons pour ce travail que la théorie Sens-Texte s'est moins intéressée aux relations paradigmatiques[7].

Lyons et Cruse Les travaux des britanniques Lyons (1977) et Cruse (1986) se rapprochent du paradigme structuraliste européen. À l'extrême celui-ci postule que le signe linguistique en lui-même n'a pas de signification et seules ses relations au sein d'un système structuré lui permettent de véhiculer du sens (De Saussure, 1995, éd. 1916)[8]. Dans ce contexte, Lyons et Cruse ont porté leur intérêt sur l'étude des relations, car considérées comme centrales dans l'étude du sens. Cruse définit le sens d'une unité au travers de ses relations contextuelles :

> The full set of normality relations which a lexical item contracts with all conceivable contexts will be referred to as its **contextual relations**.

[4] Les relations entre les unités lexicales sont exprimées par les relations entre les concepts auxquels elles sont associées.
[5] Par exemple, pour la synonymie : **Syn**(*telephone*)=*phone*
[6] Par exemple **Anti**(boy)=girl, **Anti**(boy)=man
[7] Parmi le 64 fonctions lexicales présentées dans (Mel'čuk et Wanner, 1996), un peu plus d'un tiers sont paradigmatiques.
[8] Cette position forte sera celle retrouvée dans les approches holistes et associatives.

> We shall say, then, that the meaning of a word is fully reflected in its contextual relations ; in fact, we can go further, and say that, for present purposes, the meaning of a word is constituted by its contextual relations.

Ces relations contextuelles incluent les relations syntagmatiques et paradigmatiques. Néanmoins, Cruse souligne ensuite la nécessité de ne pas s'y limiter :

> A particular lexical unit, of course, expresses its semantic identity through such relations, but its essence cannot be exhaustively characterised in terms of any determinate set of such relations.

Approches holistes et associatives Ces approches sont dites *associatives*, car elles spécifient le sens des unités lexicales par les relations qui les lient, et elles sont dites *holistes*, car elles considèrent le sens à travers l'ensemble du système : le sens d'une unité dépend de toutes les unités avec lesquelles elle est en relation (et de toutes les unités qui sont reliées à ces dernières, etc.). Nous présentons ici deux perspectives : la position de Fodor et la ressource lexicale WordNet.

Fodor Dans le travail de Fodor (1975), le sens est défini uniquement par des propositions logiques. Dans ce contexte, les relations sémantiques sont traitées comme des postulats (pour *meaning postulates*) (Fodor, 1980) et les unités seules n'ont pas de définition. Ainsi, le sens n'est donc pas dans les unités lexicales ou les concepts, mais entre eux.
Dans *Language of Thought* (1975), les arguments avancés sont essentiellement des critiques envers les approches compositionnelles. Premièrement, la représentation compositionnelle du sens est une tâche impossible à réaliser de manière parfaite. Deuxièmement, si les concepts sont composites alors un mot plus complexe devrait induire un temps plus élevé de traitement chez les locuteurs. Or les expériences de Fodor n'ont pas montré de différence significative. Par la suite, les approches compositionnelles réfuteront ces arguments (Jackendoff, 1990).

WordNet Avec WordNet, Fellbaum *et al.* (1998b) proposent une ressource lexicale où la description des unités lexicales est faite au travers des relations paradigmatiques qui les lient[9]. Deux types de relations sont distingués (Fellbaum, 1998a; Miller, 1990). Les *relations lexicales* lient les unités lexicales, comme par exemple la synonymie qui forme les *synsets* (ensembles de synonymes exprimant un concept). Les *relations conceptuelles* lient ces synsets de manière à former un arbre. Notons que le vocabulaire est celui de la linguistique : la relation de subsomption est appelée hyperonymie.
WordNet considère peu de relations, mais celles-ci sont justifiées par leur caractère cognitivement saillant pour les locuteurs, au contraire des autres approches (Fellbaum, 1998a) :

[9] WordNet ne modélise pas les relations syntagmatiques. Les noms, les verbes, les adjectifs et les adverbes sont traités séparément.

> Other lexical semanticists have undertaken careful analyses of semantic and lexical relations and proposed subtle distinctions (Cruse 1986). These distinctions are valid in the context of a semantic analysis of conceptual relations, but they do not seem to be reflected in speakers' minds, where relatively few relations are salient. Probably the largest number of relations —53— has been proposed by Mel'čuk and Zholkovsky (1988), who call them "lexical functions." Many of these include relations among morphologically related word forms.

Malgré les critiques à son encontre, WordNet a été utilisé dans un grand nombre de travaux et a été le point de départ pour l'établissement de ressources lexicales dans d'autres langues (Vossen, 1998; Sagot et Fišer, 2008).

1.1.3 Notions de terme et d'entité nommée

Les termes et les entités nommées sont deux réalités linguistiques différentes. Il est généralement partagé que ces entités textuelles fournissent des informations quant aux domaines dont elles sont extraites (Omrane *et al.*, 2011).

Terme Un terme est une unité lexicale composée d'un mot (terme simple) ou plusieurs (terme composé) et utilisée au sein d'un domaine ou d'une communauté de travail (Lerat, 2009). Dans la tradition de Wüster (1981), il est admis que le terme réfère à un concept de manière idéalement non ambiguë[10]. Cette conception stable et fixe du terme se retrouvera dans les standards internationaux ISO 704 (1987) et ISO 1087 (1990)[11].

Par la suite, cette vision uniquement référentielle a été remise en cause et le terme a été considéré comme un objet linguistique dont le comportement se rapproche de celui des autres syntagmes nominaux. Jacquemin (2001) explique :

> Despite the constant and characteristic features that give an illusion of fixedness, terms are genuine complex lexical entries - possibly polysemous, possibly structurally ambiguous - that can be modified by morphological, semantic, and syntactic transformations and integrated into the construction of novel lexical entries.

Dans ce contexte, l'importance est donnée à la forme en corpus[12] mais aussi à la dynamique du sens[13] (Poibeau, 2005a). Le sens d'un terme n'est plus une propriété intrinsèque, mais doit être reconsidéré à l'aide de la sémantique lexicale (Faber et L'Homme,

[10] Pour reprendre l'expression de Poibeau (2005a), le terme était traditionnellement considéré comme « une étiquette linguistique sur une unité du monde ».

[11] Ces standards supposent notamment que le terme soit sans synonyme et sans variante morphologique, et que sa création soit lexicalement systématique. Voir Sager (1990, p.89-90).

[12] Cet intérêt pour la forme en corpus fait écho à la position de Kilgarriff (1997) qui soutient qu'il n'y a pas de sens des mots mais uniquement des « corpus citations ».

[13] Au sens de Rastier (1996) : « Si la description statique peut convenir à certaines applications, en didactique par exemple, une description plus fine doit restituer l'aspect dynamique de la production et de l'interprétation des textes. »

2014). Par exemple, citons Meyer (2001) qui repose sur la théorie de Cruse (1986), ou Peruzzo (2014) qui utilise les cadres de Fillmore (1982) pour caractériser les termes.

La sémantique derrière les relations liant les termes a fait l'objet de recherches (Jouis, 2002), bien que, dans la pratique, les principaux utilisateurs de ressources s'avèrent être des acteurs industriels pour lesquels la consistance importe davantage que la complexité.

Entité Nommée Les entités nommées sont des unités textuelles généralement assimilées aux noms propres ainsi qu'aux valeurs numériques. Historiquement, ces unités étaient réparties parmi les trois catégories proposées aux conférences MUC-6[14] (1996) et MUC-7 (1997) : ENAMEX (personnes, organisations, etc.), TIMEX (dates, heures, etc.) et NUMEX (valeurs numériques, monnaies, etc.). Par la suite, des travaux ont proposé d'organiser leurs types sous la forme d'une hiérarchie (moyenne (Ferret *et al.*, 2001) ou grande (Hasegawa *et al.*, 2004)) ou d'une ontologie (Rizzo et Troncy, 2012).

Les campagnes ACE[15], qui succéderont aux MUC, différeront quelque peu en proposant une perspective plus sémantique. Doddington *et al.* (2004) expliquent :

> (...) the so-called "named entity" task, as defined in MUC, is to identify those words (on the page) that are names of entities. In ACE, on the other hand, the corresponding task is to identify the entity so named. This is a different task, one that is more abstract and that involves inference more explicitly in producing an answer.

Dans ce contexte, les relations entre entités nommées feront l'objet d'une tâche à part entière. Par exemple, pour ACE 2004, les types de relations suivants entre entités nommées ont été proposés : *Role* (rôles d'une personne dans une organisation), *Part* (relations partitives : membre, composant, etc.), *At* (relations de localisation : basé-à, etc.), *Near* (relations de localisation relative), *Social* (relations sociales : parent-de, frère-de, etc.).

Notons que traditionnellement, les entités nommées ont été considérées comme des ancres référentielles servant de pointeurs vers des unités externes. Kripke (1982) parle de *désignateurs rigides*[16]. De manière analogue aux termes, une remise en question de cette conception uniquement référentielle a été faite. En effet, il apparaît qu'une même entité du monde peut être désignée par plusieurs formes linguistiques (problème de synonymie) et plusieurs entités peuvent être désignées par une seule forme linguistique (problème de polysémie) (Poibeau, 2005b). Également, un recours à la sémantique lexicale est proposé dans la littérature[17].

[14] Message Understanding Conference
[15] Automatic Content Extraction
[16] « Let's call something a *rigid designator* if in every possible world it designates the same object, a *nonrigid* or *accidental designator* if that is not the case. Of course we don't require that the objects exist in all possible worlds. » (Kripke, 1982).
[17] Sur cette question, nous renvoyons au chapitre 2 du travail d'Ehrmann (2008).

1.2 Approches sur textes non structurés

Dans cette section, nous présentons les approches d'extraction de relations à partir de textes où les éléments de mise en forme ne sont pas pris en compte. Nous présentons les approches selon trois catégories :

- les approches symboliques ;
- les approches statistiques ;
- les approches hybrides.

Nous centrons notre propos sur l'extraction des relations sémantiques directement observables en corpus[18]. Dans ce contexte, les approches présentées extraient essentiellement des relations *hiérarchiques* (hyperonymie, holonymie) (Grabar et Hamon, 2004). Nous mettons de côté les relations *intra-catégorielles* (Sébillot, 2002) ainsi que les relations dites *lexicales* d'antonymie et de synonymie (Hamon et Nazarenko, 2001).

1.2.1 Approches symboliques

Les approches symboliques se fondent sur l'idée que les relations sémantiques montrent des propriétés linguistiques internes (dites structurelles) et des propriétés externes (dites contextuelles) (Nazarenko et Hamon, 2002) et qu'il est possible de les modéliser par l'écriture de règles.

Approches structurelles L'étude des propriétés linguistiques internes entre unités lexicales permet généralement d'obtenir de bons résultats. Citons notamment les méthodes qui inspectent les dépendances syntaxiques au sein du terme comme le fait LEXTER (Bourigault, 1994) pour aider l'expert dans la création d'un réseau terminologique. Citons également les travaux reposant sur la variation morphologique tels que les travaux de Jacquemin avec l'outil FASTR (1994; 1996; 1997) ainsi que les travaux de Grabar (1999) pour le domaine médical. Dans les domaines contrôlés, Grefenstette (2015) a montré qu'un simple test d'inclusion lexicale s'avérait productif : « This heuristic was unexpectedly productive in the chemical domain where many hypernym pairs were similar to: *ginsenoside mc* as a type of *ginsenoside*. »

Approches contextuelles L'étude des propriétés linguistiques externes s'est surtout faite au travers de la mise en place de patrons lexico-syntaxiques. Les travaux pionniers dans ce domaine sont ceux de Hearst (1992). La méthode proposée initialement pour la relation d'hyperonymie sera ensuite déclinée par d'autres relations telles que l'holonymie (Berland et Charniak, 1999) ou les relations causales (Garcia, 1997). L'intuition

[18] L'accès direct au corpus permet la définition de la nature de la relation, contrairement aux approches distributionnelles (et plus généralement statistiques globales (Nazarenko et Hamon, 2002)) qui extraient des classes d'unités lexicales mais où la relation est généralement sous-catégorisée (Séguéla, 2001). Grefenstette (1994) parle de similarité et utilise les distributions de contextes lexico-syntaxiques des termes à lier.

derrière les patrons lexico-syntaxiques réside dans le fait que pour une relation sémantique donnée, il existe des régularités prévisibles qui expriment cette relation. Hearst (1992) explique :

> We identify a set of lexico-syntactic patterns that are easily recognizable, that occur frequently and across text genre boundaries, and that indisputably indicate the lexical relation of interest.

L'appellation de patron lexico-syntaxique n'est pas partagée uniformément. Aussenac et Séguéla (2000; 2001) parlent de « formules linguistiques » qui une fois implémentées sont appelées « schémas ». Pour Meyer (2001), les « patrons de connaissance » (pour *Knowledge Patterns*) permettent la découverte de « contextes riches en connaissances » (pour *Knowledge-Rich Contexts*). Plus récemment, Arnold et Rham (2014) ont utilisé l'expression de « patrons sémantiques » (pour *semantic patterns*).

Pour trouver des patrons exprimant des relations d'intérêt, Auger et Barrière (2008) répertorient deux approches : (i) l'approche onomasiologique qui démarre d'une relation donnée et cherche ensuite à identifier les marqueurs qui l'expriment, et (ii) l'approche sémasiologique qui cherche à identifier les relations exprimées par des marqueurs fixes.

Hearst (1992) a proposé une approche onomasiologique appelée amorçage (*bootstrapping*). Celle-ci, ainsi que les travaux qu'elle a influencés seront détaillés en section 1.2.3.

Méthodologies pour les approches contextuelles L'avantage des approches par patrons réside dans la possibilité d'adapter finement les patrons pour une relation donnée. La qualité de la relation est généralement privilégiée à la quantité (Séguéla, 2001). Ainsi, les résultats montrent généralement une bonne précision, mais un rappel bas. Pour reprendre Cimiano (2005) cité par Auger et Barrière (2008) :

> The approaches of Hearst and others are characterized by a (relatively) high precision in the sense that the quality of the learned relations is high. However, these approaches suffer from a very low recall which is due to the fact that the patterns are very rare in corpora.

En outre, un coût généralement élevé est associé à la définition manuelle (ou semi-manuelle) des patrons et l'étape de création est généralement à refaire pour un nouveau domaine (Jacques et Aussenac-Gilles, 2006). Pour remédier à ce coût, plusieurs méthodologies visant la génération et la réutilisation des patrons ont été proposées.

Jouis (1997), et dans la suite Le Lepriol (2001), ont proposé l'outil SEEK permettant l'identification et la gestion des relations sémantiques extraites. Le modèle linguistique implémenté est celui de la Grammaire Applicative et Cognitive (Desclés, 1990). Dans ce cadre, l'acquisition des relations se fonde sur l'*exploration contextuelle* : à l'aide de listes de marqueurs, la première étape marque les relations sémantiques, ensuite il est possible d'appliquer des patrons lexico-syntaxiques pour trouver les arguments (Desclés, 2006).

Condamines et Rebeyrolle (1997) proposent une méthode pour la constitution de bases de connaissances terminologiques. La problématique de la dépendance au corpus est soulevée [19] et les auteurs préconisent (i) une meilleure sémantisation des éléments extraits au sein des outils utilisés, (ii) une plus grande souplesse d'utilisation dans ces outils et (iii) la nécessité d'un processus interactif entre le texte et les patrons extraits.

Morin (1998; 1999) a proposé le système PROMÉTHÉE. Celui-ci utilise des patrons pour extraire des relations sémantiques, mais s'appuie sur une analyse distributionnelle pour chaque nouveau domaine. Ainsi, il est possible de faire émerger rapidement des marqueurs récurrents. Pour évaluer le système, Morin a proposé une étude sur la relation d'hyperonymie.

Séguéla (1999; 2001) présente la méthode CAMELEON qui permet d'acquérir rapidement des connaissances d'un domaine, et d'assurer l'adaptation des marqueurs extraits sur d'autres domaines. La méthode proposée est similaire à celle de Hearst.

Dans la suite des approches symboliques contextuelles, les approches statistiques ont été proposées pour à la fois augmenter la couverture, mais également diminuer le coût lié à la création et à la maintenance des patrons.

1.2.2 Approches statistiques

Dans cette section, nous décrivons les approches statistiques pour l'extraction de relations. Ces approches se fondent sur l'hypothèse qu'il est possible d'apprendre les marqueurs contextuels à partir de régularités et de les généraliser sur de nouvelles données. Dans ce cadre, le problème d'extraction de relations est généralement considéré comme un problème de classification binaire.

Les approches statistiques se divisent en deux catégories : les approches supervisées et celles qui ne le sont pas (approches semi-supervisées et non-supervisées). Les frontières entre ces catégories ne sont pas précises et un continuum doit être considéré. Notons également que cette classification varie d'une étude à l'autre selon les propriétés mises en avant par les auteurs[20].

Approches supervisées Les approches supervisées permettent d'apprendre les marqueurs contextuels à partir d'un ensemble de relations déjà extraites et annotées manuellement appelé ensemble d'apprentissage. Nous reprenons ici la classification faite par Bach et Badskar (2007) où deux familles de méthodes sont considérées : les classifieurs à base de traits et les méthodes à noyaux.

[19] Voir les travaux (Condamines, 2003, 2008) pour une réflexion plus générale sur le lien entre corpus et acquisition de connaissances.

[20] Par exemple, les approches décrites comme semi-supervisées par Cartier (2015) correspondent aux approches que nous décrivons comme supervisées.

Classifieurs à base de traits Le terme de classifieurs à base de traits désigne la catégorie des classifieurs d'apprentissage statistique où les observations doivent être encodées manuellement sous la forme de traits, c'est-à-dire sous la forme de paires attribut-valeur décrivant les observations. L'intérêt pour l'extraction de relations à partir de ce type de classifieurs a été grandissant au début des années 2000. Dans la tâche d'extraction de relations de la campagne MUC-7 (1998), seul un système statistique est proposé (SIFT, (Miller *et al.*, 1998)). Avec l'essor des campagnes ACE, le nombre de travaux avec des classifieurs a été grandissant. Il sera question de « ACE-style algorithms »[21]. Citons par exemple Kambhlatla (2004) qui propose l'utilisation d'une régression logistique multinomiale et GuoDonc *et al.* (2005) qui utilisent une Machine à Vecteurs de Support (pour *Support Vector Machine* (SVM)). Rosario et Hearst (2004) ont présenté une comparaison entre un modèle graphique génératif et un réseau de neurones discriminant dans le domaine de la bioscience[22].

Méthodes à noyaux Les méthodes à noyaux (Vapnik, 1998) consistent à transformer l'espace de traits original en un autre de plus grande dimension où il aura plus de chance d'être séparé linéairement[23]. Ces méthodes ont montré leur avantage dans des tâches telles que le parsing syntaxique (Collins et Duffy, 2001) ou la catégorisation de textes (Joachims, 2002). Une de leurs propriétés intéressantes réside dans le fait qu'il n'est pas nécessaire de définir manuellement un grand nombre de traits. Les exemples d'apprentissage peuvent être utilisés dans leur représentation presque originelle et la fonction noyau permet de déterminer la classe d'une nouvelle observation. Pour l'extraction de relations, l'un des travaux pionniers est celui de Zelenko *et al.* (2003) qui propose une méthode à noyaux sur des arbres de constituants syntaxiques. De manière comparable aux classifieurs à base de traits, de nombreuses études ont été menées sur les tâches ACE. Citons Culotta et Soresen (2004) et Bunescu et Mooney (2005) qui utilisent des arbres de dépendances syntaxiques. Également, citons Zhao et Grishman (2005) qui associent un noyau à chaque niveau : tokens, arbre de constituants et arbre de dépendances.

Néanmoins, les approches supervisées reposent sur l'annotation d'un ensemble d'apprentissage, ce qui est coûteux (Mintz *et al.*, 2009). Les approches semi-supervisées et non-supervisées visent à s'affranchir de cette contrainte.

[21] Expression reprise de Mintz *et al.* (2009). Ces auteurs déclareront également « Modern models of relation extraction for tasks like ACE are based on supervised learning of relations from small hand-labeled corpora. »

[22] La comparaison génératif - discriminant est une comparaison courante en apprentissage (Ng et Jordan, 2002; Sutton et McCallum, 2006). Nous reviendrons sur celle-ci dans le chapitre 5 pour justifier le choix de nos algorithmes.

[23] L'intuition est la suivante : « A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated. » (Kim *et al.*, 2005)

Approches semi-supervisées et non-supervisées Les approches semi-supervisées et non-supervisées permettent d'apprendre sans que cela ne nécessite une intervention humaine lourde dans la phase d'annotation. Deux catégories sont présentées ici : la supervision distante et l'extraction ouverte de relations.

Supervision distante Initialement proposée par Craven *et al.* (1999) et ensuite popularisée par Snow *et al.* (2004) et Mintz *et al.* (2009), la supervision distante est une méthode aujourd'hui largement répandue. Elle repose sur la combinaison d'une ressource (p. ex. WordNet, Freebase, etc.) avec un corpus de très large échelle. Les relations trouvées entre deux entités textuelles[24] de la ressource sont associées aux phrases où ces entités apparaissent. Ces phrases servent alors d'exemples d'entraînement positifs pour un classifieur discriminant[25]. Néanmoins, cette hypothèse n'est pas toujours vérifiée[26]. Riedel *et al.* (2010) proposent l'introduction d'un modèle graphique pour modéliser cette incertitude. Les travaux actuels en supervision distante continuent dans le sens d'une meilleure sélection des phrases pour l'entraînement (Hoffmann *et al.*, 2011; Xu *et al.*, 2013).

Extraction ouverte de relations Contrairement aux travaux présentés précédemment, l'extraction ouverte de relations propose d'extraire des connaissances sans pour autant que les relations recherchées soient définies *a priori*. Hasegawa *et al.* (2004) présentent une méthode de partitionnement de données (pour *clustering*) qui repose sur l'idée que les paires d'entités classées dans les mêmes groupes partagent la même relation. Le critère de regroupement est alors fait sur le partage des contextes. Banko *et al.* (2007) proposent TextRunner[27] qui permet une extraction à large échelle. Le procédé est le suivant : (i) entraîner un classifieur auto-supervisé sur un petit corpus analysé syntaxiquement, (ii) extraire les relations candidates dans un corpus large, (iii) assigner une probabilité à chaque relation extraite en fonction de sa fréquence. La tendance actuelle se tourne vers l'intégration de modèles distributionnels (Yao *et al.*, 2012; Akbik *et al.*, 2012).

Toutefois, les méthodes statistiques reposent toutes sur l'hypothèse que la quantité donnée disponible permette l'apprentissage de régularités utiles pour extraire les relations. De facto, les phénomènes rares ou avec une grande variabilité sont difficilement pris en compte. Les approches hybrides permettent de combiner l'intégration de connaissances humaines avec des outils statistiques.

[24] Dans ce contexte, la distinction terme et entités nommées est très rarement faite.
[25] Les exemples négatifs sont générés en un nombre équivalent aux exemples positifs avec des procédés simples.
[26] Par exemple, la relation joue_avec(Maxime,chat) ne peut pas être associée à la phrase *Maxime est allergique au chat.*
[27] Renommé ReVerb - http://reverb.cs.washington.edu/

1.2.3 Approches hybrides

Les approches hybrides se fondent sur la définition de patrons lexico-syntaxiques ainsi que sur l'utilisation de statistiques. Notons que certains systèmes présentés précédemment, tels que celui de Morin (1999), pourraient aussi rentrer dans cette section. Nous centrons notre propos sur la technique d'amorçage (pour *bootstrapping*).

Amorçage La méthode introduite par Hearst (1992) est la suivante : à partir d'un petit ensemble de couples d'entités textuelles (appelé *seed*) il est possible, en les projetant sur un corpus, de découvrir de nouveaux patrons lexico-syntaxiques. La qualité de ces derniers est évaluée et, une fois validés, ceux-ci sont utilisés pour trouver de nouvelles entités textuelles. En 1992, l'ensemble du procédé est manuel. Par la suite, Hearst (1998) propose une version automatisée. Cette méthodologie incrémentale influença plusieurs travaux. Leur évolution sera liée à celle du Web et de l'explosion des données.

Brin (1998) propose la méthode DIPRE et présente une expérience pour la relation *auteur-de*[28]. Cette méthode reprend le principe de Hearst, mais évalue de manière interdépendante les patrons et les couples extraits. Le système Snowball (Agichtein et Gravano, 2000) étendra cette méthode. Dans KnowItAll, Etzioni *et al.* (2004) démarrent le processus d'amorçage en définissant d'abord manuellement une série de patrons corrects pour une liste de classes (p. ex. ville, scientifique, etc.) et, ensuite, ils trouvent des instances de ces classes sur le Web. Cimiano *et al.* (2004) proposent le système PANKOW. Dans ce travail, le problème lié à la rareté des patrons sur de petits corpus est contourné en utilisant les documents du Web. Les patrons initiaux sont ceux de Hearst (1992), ensuite, le système découvre cycliquement de nouveaux patrons en faisant remonter des couples trouvés au travers de l'API Google. Alfonseca *et al.* (2006) proposent une métrique de *spécificité* pour évaluer les patrons. Un patron très spécifique est un patron qui exprime clairement une relation et qui n'est pas ambigu. Enfin, contrairement aux approches précédentes qui prennent n mots autour des couples projetés, dans Espresso (Pantel et Pennacchiotti, 2006) les auteurs évaluent des patrons en utilisant l'ensemble de la phrase.

L'acquisition automatique de nouveaux patrons lexico-syntaxiques permet de réduire le coût lié à l'intervention humaine dans ce type de système. Toutefois, un problème de glissement sémantique (pour *semantic drifting*) peut survenir lors des phases d'expansion : un patron de qualité moindre, une fois validé, va entraîner de manière cumulative de mauvais couples au fur et à mesure des cycles (Cartier, 2015). Les recherches actuelles pour régler ce type de problème se tournent vers l'intégration d'une mesure de similarité entre contextes syntaxiques des patrons (Zhang *et al.*, 2014; Lambrou-Latreille, 2015).

[28] Bien que le travail de Brin lie des entités nommées, ce travail est largement cité dans les revues de terminologies (Auger et Barrière, 2008). Ceci illustre en partie le flou qui peut exister dans la littérature en extraction de relations à propos de la nature des entités textuelles liées.

1.3 Approches sur textes structurés

Les approches sur des textes structurés visent à extraire des relations sémantiques en utilisant les propriétés de formatage et de mise en forme des documents qu'elles traitent. Deux catégories d'approches sont distinguées :

- les approches exploitant des formatages prédéfinis ;
- les approche exploitant des documents à balises.

1.3.1 Approches exploitant des formatages prédéfinis

Cette section présente les approches qui exploitent le formatage spécifique des documents ou des parties de ceux-ci pour lesquelles la sémantique est relativement stable.

Documents formatés Traditionnellement, les dictionnaires électroniques (pour *machine readable dictionary*) furent utilisés comme source pour acquérir des relations (Auger et Barrière, 2008). Généralement issus du format papier, ces dictionnaires sont difficiles à exploiter. Le formatage des informations pertinentes n'est pas systématique, et le langage utilisé reste du langage naturel (par opposition à un langage formel). Nous présentons ici quelques travaux qui ont proposé des solutions pour contourner ces difficultés. Véronis et Ide (1991) montrent que l'exploitation de multiples dictionnaires permet d'améliorer la précision :

> It therefore appears that even if individual dictionaries are an unreliable source of semantic information, multiple dictionaries can play an important role in building large lexical-semantic databases.

MindNet (Richardson *et al.*, 1998) est une ressource lexicale construite à partir des définitions et des phrases d'exemples issues de deux dictionnaires électroniques. Sur les arbres produits en sortie du parseur de Microsoft Word 97, les auteurs appliquent des patrons lexico-syntaxiques. Les structures ainsi extraites sont introduites dans MindNet et servent ensuite d'amorce pour des traitements ultérieurs (p. ex. inférence, désambiguïsation, etc.). Jannink (1999) propose d'extraire un thésaurus à partir du dictionnaire Webster en modélisant le problème sous la forme d'un graphe et en extrayant les relations selon un algorithme comparable au PageRank. Citons aussi les travaux de Wilks (1993), ou les travaux de Rigau (1998). Néanmoins, l'avènement du Web a amené les chercheurs à se tourner vers des ressources de natures différentes (Kilgarriff et Grefenstette, 2003).

Navarro *et al.* (2009) et par la suite Sajous *et al.* (2011) utilisent Wikitionary[29] afin d'extraire des relations de synonymie. La méthode proposée est une marche aléatoire calculant la similarité entre les entrées lexicales. Barque *et al.* (2010) proposent d'utiliser le Trésor de la Langue Française informatisé (TLFi)[30] pour extraire des relations

[29] Wikitionary est un dictionnaire en ligne collaboratif qui se positionne comme un projet satellite à l'encyclopédie Wikipédia. Zesch (2008) parle de « lexical companion ».

[30] Le Trésor de Langue Française informatisé est une ressource lexicographique maintenue par le laboratoire ATILF.

d'hyperonymie. Plus récemment, Cartier (2015) propose d'extraire des relations à partir du Wikitionary et du TLFi, mais aussi de l'encyclopédie Wikipédia. La méthode proposée combine patrons lexico-syntaxiques et l'outil statistique SDMC (Béchet *et al.*, 2013) pour repérer les contextes pertinents.

Structures textuelles localisées Il existe des structures pour lesquelles la sémantique est relativement stable et qui peuvent être exploitées pour l'acquisition de relations.

Les travaux de Navigli *et al.* (2007; 2008) proposent d'exploiter les gloses. Les gloses sont des définitions décrivant des termes pour un domaine donné. Généralement, les premières phrases contiennent l'expression de relations à caractère définitoire (p. ex. hyperonymie). Ces travaux rejoignent ceux sur la définition (Rebeyrolle et Tanguy, 2000; Malaisé *et al.*, 2004).

Chernov *et al.* (2006) proposent une méthode pour détecter des relations entre les catégories de Wikipédia. L'hypothèse avancée est que les catégories en relation correspondent à un grand nombre de pages connectées par des liens hypertextes. La métrique utilisée compte les liens hypertextes entrants et sortants. Les résultats montrent qu'il est possible d'identifier des liens mais ceux-ci restent sous-spécifiés.

Suchanek *et al.* (2007) extraient les relations entre les catégories et les titres des pages auxquelles elles sont associées (p. ex. type(Einstein, physicist)). La méthode utilise des heuristiques pour chaque type de relations (p. ex. *Type*, *subClassOf*, etc.). WordNet est exploité conjointement pour désambiguïser. Les résultats ont conduit à la ressource YAGO. Hoffart *et al.* (2013) proposeront ensuite YAGO2. Dans cette extension, un intérêt est porté aux informations temporelles et spatiales, et d'autres structures textuelles sont exploitées (infoboxes, tables, etc.).

Auer *et al.* (2007) exploitent les infoboxes de Wikipédia. Les infoboxes sont des tables qui décrivent l'entité d'une page à l'aide de paires attribut-valeur. Par exemple, la page Innsbruck contient la paire attribut-valeur Country-Austria. Avec des heuristiques, les auteurs montrent qu'il est possible d'extraire des relations (p. ex. hasCountry(Innsbruck, Austria). Les relations ainsi extraites servent de noyau à la ressource DBpedia (2007).

Cafarella *et al.* (2008) proposent le système WebTables qui exploite les tables HTML. La méthode repose sur le filtrage des tables qui ne comportent pas de relations (c.-à-d. celles propres aux menus, aux formulaires, etc.), ainsi que sur l'acquisition de la nature de la relation au travers des intitulés des colonnes. Les auteurs estiment que sur les 14,1 milliards de tables dans leur corpus[31], seul 1,1% contient des relations sémantiques exploitables. Ces travaux ont été étendus par Venetis *et al.* (2011) et les relations extraites ont été utilisées pour le Knowledge Vault (Dong *et al.*, 2014).

1.3.2 Approches sur des textes à balises

Dans cette section, nous présentons les approches qui exploitent les balises des textes pour l'extraction de relations. Nous regroupons ces approches selon le type de langage à balises qu'elles traitent : le format HTML, le format XML et le format WikiText.

[31] Il s'agit de l'index Google.

Format HTML Le format HTML est l'un des langages à balises les plus immédiats pour la structuration des documents sur le web.

Un des travaux pionniers est celui de Shinzato et Torisawa (2004a). Les auteurs proposent d'extraire des relations d'hyperonymie en utilisant les listes extraites de pages HTML en japonais. L'hypothèse est que les entités[32] exprimées dans une même liste ont tendance à partager le même hyperonyme commun. La méthode est la suivante : (i) les entités partageant les mêmes chemins HTML (p. ex.) sont regroupées en groupes d'hyponymes candidats, (ii) un hyperonyme candidat est sélectionné dans un large corpus avec des métriques, (iii) les hyperonymes et les groupes d'hyponymes sont appariés en évaluant la similarité de leur vecteur de co-occurrences.

Dans la suite, Shinzato et Torisawa (2004b) cherchent à extraire l'hyperonyme directement à partir des éléments encadrant la liste. Le procédé est le suivant : pour un ensemble préalable d'hyperonymes candidats, les auteurs téléchargent un corpus web où ceux-ci apparaissent dans des configurations de titre[33], ensuite ces hyperonymes sont appariés aux groupes d'hyponymes extraits d'une manière semblable à la première méthode.

Les travaux de Shinzato et Torisawa ont ouvert la voie à de nombreux autres travaux. Yoshinaga et Torisawa (2007) visent à extraire de manière non-supervisée des paires attribut-valeur pour une entité donnée (p. ex. Ben-Hur avec la paire <director, W. Wyler>). L'hypothèse est que les sites web décrivent des classes identiques d'entités en utilisant des attributs communs, et ces attributs sont généralement mis en avant dans la structure HTML[34]. La méthode présente deux étapes : (i) extraire grâce à des mesures statistiques les attributs pour une classe donnée (p. ex. la classe movie et les attributs director, runtime, etc.) et (ii) extraire les valeurs associées grâce à des patrons. Notons ici que les auteurs postulent que les valeurs d'un attribut le suivent directement. La chaîne de caractères comprise entre un attribut et l'attribut qui lui succède est alors considérée comme la valeur recherchée. La problématique de la segmentation n'est pas évoquée[35].

Le travail de Ravi et Paşca (2008) vise de manière analogue à extraire des paires attribut-valeur en utilisant la structure des documents web. Les étapes sont similaires (extraire les attributs, extraire les valeurs), mais les auteurs introduisent des fonctions de score améliorant la précision. Notons que la problématique de la segmentation des entités n'est pas non plus évoquée.

Bounhas et Slimani (2010) proposent un travail exploitant la structure hiérarchique des documents HTML. Toutefois, leur contribution porte davantage sur l'indexation hiérarchique des termes, rejoignant la littérature combinant indexation et structure (Kruschwitz, 2001; Zargayouna, 2004; Faessel, 2011), plutôt que sur l'extraction de relations pour laquelle aucun indice linguistique n'est pris en compte.

[32] Les auteurs utilisent le terme d'« expression ».

[33] Les patrons suivants, traduits du japonais vers l'anglais par les auteurs, sont utilisés : *table of X*, *guide to X*, *category of X*, *list of X*, *vote to X*, *menu of X*, *ranking of X*.

[34] « (...) the attributes are likely to be emhasized in visually distinguishable ways through HTML tags and symbolic decorations » (Yoshinaga et Torisawa, 2007)

[35] Par exemple, la paire attribut-valeur <starring, "Charlton Heston, Jack Hawkins"> est considérée comme recevable.

Format XML L'extraction de connaissances en exploitant la structure exprimée au format XML est une tâche relativement récente. L'un des premiers workshops dédiés à cette tâche est le *Knowledge Discovery from XML Documents* (Nayak et Zaki, 2006).

Dans ce workshop, Brunzel et Spiliopoulou (2006) présentent le système XTREEM (Xhtml TREE Mining) dédiés aux documents XHTML[36] et visant l'extraction de relations horizontales entre termes (p. ex. synonymie, co-hyponymie, etc.). Les auteurs postulent que les termes liés apparaissent dans des configurations relativement identiques au sein de la structure du document. Cette intuition est implémentée dans la technique Group-By-Path qui regroupe les termes partageant les mêmes chemins XML[37]. Cette technique partage des propriétés avec l'approche de Shinzato et Torisawa (2004a).

D'autres travaux ont proposé des mécanismes pour extraire les relations hiérarchiques. Role et Rousse (2006) exploitent simultanément la structure et le contenu de documents pour élaborer une ontologie de la flore tropicale. Une caractéristique de leur corpus (Collection *Flore du Cameroun*) est sa structuration régulière malgré une publication étalée (1963-2001). Par conséquent, les auteurs proposent de traduire la hiérarchie induite par la description des espèces de fleurs en une taxonomie. La méthode repose sur la conversion en XML des pages à l'aide d'un logiciel d'OCR, et ensuite sur un découpage suivant la structuration. Les résultats montrent que par des traitements simples et un corpus structuré, il est possible d'obtenir l'ossature d'une ontologie.

Aussenac-Gilles et Kamel (2009) proposent d'utiliser la sémantique associée aux balises. Le corpus analysé est un ensemble de spécifications de base de données. Dans ce contexte, le schéma de la base de données est reflété dans la structure du document et il est possible de dériver des relations hiérarchiques à partir des dépendances entre balises. Toutefois, il est nécessaire qu'un expert définisse *a priori* la sémantique à donner à ces dépendances. Les auteurs explicitent :

> The identification of all correspondences between XML specifications and ontology elements did not raise any major difficulty as long as tag labels convey their own semantics and relations can be easily identified with some common-sense knowledge.

Ce travail sera par la suite étendu par Laignelet *et al.* (2011).

Format WikiText Le format WikiText est le langage dont l'interprétation offre le rendu HTML des pages de Wikipédia (ainsi que des projets satellites). À l'heure actuelle, seul le logiciel MediaWiki[38] supporte pleinement la syntaxe du format WikiText[39]. Néanmoins, l'expression de la hiérarchie est relativement stable et plusieurs travaux ont proposé de l'exploiter pour extraire des relations sémantiques.

[36] Le format XHTML est l'adaptation du standard HTML aux contraintes syntaxiques du format XML.
[37] Nous renvoyons au chapitre de Brunzel dans l'ouvrage de Buitelaar et Cimiano (Brunzel, 2008) pour une description plus complète.
[38] `https://www.mediawiki.org/wiki/MediaWiki`
[39] Pointé par Dohrn et Riehle (2011), le format WikiText n'a pas de grammaire définie et la création de parseurs robustes est une tâche difficile.

Sumida et Torisawa (2008) proposent d'extraire des relations d'hyperonymie à partir des structures hiérarchiques de Wikipédia. De manière analogue au travail de Shinzato et Torisawa (2004b), l'hypothèse est faite qu'un hyperonyme a tendance à être localisé à un niveau directement supérieur à son hyponyme dans la structure du document. La méthode d'acquisition est constituée de trois étapes : (i) chaque contenu textuel mis en forme (*titre*, *item*, *terme*) est projeté sur son subordonné direct afin de former une paire hyperonyme-hyponyme candidate, (ii) une série de patrons[40] est utilisée pour filtrer les paires candidates, et (iii) les paires non reconnues par les patrons sont fournies à un classifieur discriminant.

Cette méthode fut améliorée par Sumida *et al.* (2008) en généralisant le procédé de création des paires candidates. Dans la suite, Oh *et al.* (2009; 2010) ont présenté une méthode de co-apprentissage qui exploite les données issues de langues différentes ou de textes différemment structurés (structuré et non-structuré). Yamada *et al.* (2009; 2011) ont proposé de croiser les relations extraites de la structure des articles Wikipédia avec celles extraites des modèles distributionnels.

1.4 Discussion

Dans ce chapitre, nous avons examiné théoriquement l'objet de ce travail, la relation sémantique, et nous avons présenté quelques approches pour son extraction dans des textes en langage naturel. Deux familles d'approches ont été présentées : les approches utilisant le contenu textuel et les approches exploitant le contenu textuel avec des éléments de mise en forme.

La phase d'interprétation au cours de laquelle le lien est fait entre les termes et les concepts est rarement effectuée dans la littérature en extraction de relations[41]. Ceci s'explique notamment par le fait que les systèmes TAL utilisent généralement le versant lexical des ressources (Nazarenko, 2005) et la distinction lexical-conceptuel n'est plus alors pertinente. Murphy explicite (2003) cette position :

> Since a computer typically interacts with the world through some kind of language input (...), it might be more economical for semantic knowledge to be represented as lexical knowledge, rather than trying to maintain the linguistic/conceptual division that might be required for a model of human cognition.

Les liens entre modélisation linguistique et modélisation conceptuelle ont été le sujet de plusieurs travaux (Woods, 1975; Lenci, 2001; Hirst, 2009).

[40] Ceux-ci portent sur l'hyperonyme : *list of X*, *typical X*, *notable X*, etc.

[41] Notons néanmoins Laignelet *et al.* (Laignelet *et al.*, 2011) qui expriment explicitement leur choix : « Bien que les relations trouvées par les patrons soient des relations lexicales, nous les représentons au niveau conceptuel afin de réaliser une évaluation sur les résultats produits. Nous sommes conscients de gommer ainsi une phase délicate de l'interprétation, qui consiste à décider si un terme doit être associé à un concept existant ou donner lieu à la naissance d'un concept ou d'une instance de concept. »

La distinction entre les termes et les entités nommées n'est pas toujours nette et certains travaux choisissent (explicitement ou non) de ne pas la faire. Par exemple, Nobata, Collier et Tsujii (2000) considèrent les noms de protéines comme des entités nommées. Le modèle d'entités nommées du projet Quaero (Grouin *et al.*, 2011) étend également la définition d'entités nommées à des éléments qui ne sont pas obligatoirement des noms propres. Cela sera notamment repris par Dutrey *et al.* (2012) dans leur modélisation des entités nommées du domaine EDF (p. ex. *facture rectificative*, *agent de relève*, etc.).

Dans le domaine de l'extraction de relations, cela a des conséquences sur la nature et la portée des relations recherchées. Sumida et Torisawa (2008) précisent :

> Linguistic literature, e.g. (A.Cruse, 1998), distinguishes hyponymy relations, such as "national university" and "university", and concept-instance relations, such as "Tokyo University" and "university". However, we regard concept-instance relations as a part of hyponymy relations in this paper because we think the distinction is not crucial for many NLP applications.

De nombreux autres travaux en extraction de relations font un même choix, mais sans toutefois le reconnaître explicitement (Hearst, 1992; Snow *et al.*, 2004).

Les approches qui exploitent les éléments structurels sont plus rarement évoquées dans la littérature. Comme souligné par Role et Rousse (2006) la contribution de la structure est généralement sous-estimée, voire même considérée comme un obstacle. Dans ce contexte, l'enlèvement des balises structurelles est jugé comme une étape préalable à des traitements ultérieurs.

Nous nous situons dans la catégorie des approches d'extraction de relations à partir de textes structurés. Toutefois, notre travail se positionne orthogonalement en considérant des textes où la structure n'est pas obligatoirement décrite par un formatage prédéfini ou un ensemble de balises. Ce choix nécessite une abstraction des éléments concrets de mise en forme (p. ex. retraits, retours à la ligne, etc.). Ce sujet sera celui du chapitre 2.

Chapitre 2

Structure de document

Sommaire

Dans ce chapitre, nous portons notre attention sur les modèles théoriques et les approches pratiques en lien avec la structure de document. Les points de vue de deux communautés scientifiques sont décrits : ceux de la communauté de Traitement Automatique du Langage (TAL) et ceux de la communauté d'Analyse du Document (AD).

En TAL, plusieurs modèles théoriques de structuration du document ont été proposés dans le cadre de la génération de textes. Ces modèles visent la définition de la structure logique pour mieux traiter son interaction avec la structure rhétorique. En AD, les approches ignorent l'aspect discursif et adoptent la vision classique séparant la structure visuelle et la structure logique, sans que cette dernière ne soit clairement définie. Dans ce contexte, l'intérêt est essentiellement porté sur la découverte d'indices révélateurs d'unités définies de manière *ad hoc* en fonction de la tâche et du type de document.

Dans ce chapitre, nous développons les méthodes de deux courants. Dans un premier temps, nous présentons les modèles théoriques utilisés en TAL. Dans un second temps, nous présentons les approches empiriques en AD. Enfin, nous présentons le traitement de la structure logique dans quelques formats de documents.

2.1 Modèles théoriques de structure de document

En Traitement Automatique du Langage, trois structures du document sont généralement considérées : la *structure visuelle*, la *structure logique* et la *structure discursive*. Les frontières entre ces structures ne sont pas nettement établies. Toutefois, il est admis que ces structures s'échelonnent graduellement dans la compréhension et entretiennent des relations complexes d'interdépendance. Par exemple : la mise en forme spatiale d'un texte a des répercussions sur l'interprétation de sa structure logique (Virbel *et al.*, 2005), tandis qu'une relation de coordination entre deux items d'une structure hiérarchique implique une relation rhétorique spécifique (Vergez-Couret *et al.*, 2011). Dans cette section, nous présentons les principaux modèles théoriques proposés en TAL.

2.1.1 Modèle de Power *et al.* (2003)

Le modèle de Power *et al.* (2003) propose la *Document Structure*. Celle-ci est une structure abstraite et séparée de la description physique du document. Elle se positionne entre la structure physique (pour *physical representation*) et la structure rhétorique (pour *rhetorical structure*). Son rôle est de décrire l'organisation du document décomposé en constituants graphiques tels que les sections, les paragraphes, les figures, les listes, les phrases ainsi que les éléments intra-phrastiques (p. ex. marques d'emphase). L'hypothèse avancée est que la présentation graphique interagit avec le texte et son sens :

> The overlay of graphics on text is in many ways equivalent to the overlay of prosody on speech. Just as all speech has prosody (even if it is a monotone), so too do all texts have layout (even if it is simple wrapped format, in a single face and font, and makes rudimentary use of white space). And just as prosody undoubtedly contributes to the meaning of utterances, so too does a text's graphical presentation contribute to its meaning.

Dans la suite de la section, nous présentons la grammaire de Nunberg qui est le socle de la *Document Structure*, ensuite, nous montrons la syntaxe utilisée et, enfin, nous développons son lien avec l'organisation rhétorique.

Grammaire de Nunberg La *Document Structure* trouve son origine dans le travail de Nunberg. Dans *The Linguistics of Punctuation* (1990), celui-ci distingue deux grammaires :

- une grammaire relative aux propriétés linguistiques (*lexical grammar*)[1] ;
- une grammaire relative aux propriétés visuelles (*text grammar*).

La grammaire relative aux propriétés visuelles distingue la dimension concrète des textes, exprimée au travers d'indicateurs textuels (*text-category indicators*), et la dimension abstraite des textes, décomposée en catégories textuelles (*text-categories*).

[1] Celle-ci correspond à la grammaire traditionnelle.

Le terme indicateur textuel peut désigner des caractères particuliers (p. ex. marques standards de ponctuation), les alternances de fontes (p. ex. taille, police, casse, etc.) ou l'utilisation de l'élément vide[2] (p. ex. alinéa, tabulation, marge, etc.). Ces indicateurs permettent de marquer les éléments issus des catégories de trois manières : (i) en délimitant le début ou la fin de ces éléments (*delimiter*), (ii) en séparant deux éléments de même catégorie (*separator*) et (iii) en distinguant un élément d'une catégorie donnée au sein de son contexte visuel (*distinguisher*) (Nunberg, 1990, p. 17, p. 52-53).

Les catégories textuelles sont les classes de constituants graphiques telles que le paragraphe, la phrase (*text-sentence*), la clause (*text-clause*), etc. La syntaxe introduite par la grammaire permet d'organiser hiérarchiquement ces éléments les uns par rapport aux autres. Dans la tradition générativiste, les règles de cette syntaxe formulées par Nunberg sont des règles de réécriture. Par exemple, la règle de réécriture suivante :

$$S_t \rightarrow C_t^+$$

implique que la phrase (*text-sentence*) S_t est composée par une séquence de une ou plusieurs clauses (*text-clauses*) C_t. Cette règle distingue la phrase visuelle de la notion traditionnelle et syntaxique de phrase (Power, 2000)[3].

La distinction entre abstrait (catégories) et concret (indicateurs) permet de séparer la structure du texte (*text-structure*) de sa réalisation physique. Ne pas opérer cette distinction présente des inconvénients. Par exemple, si pour un éditeur donné, un paragraphe débute avec un retour à la ligne et un alinéa, un autre éditeur pourra cependant préférer deux retours à la ligne sans alinéa. Ainsi, si la structure logique est réalisée par les indicateurs, il n'y a néanmoins pas d'équivalence assurée entre les deux.

Syntaxe de la *Document Structure* Dans le système de génération de textes ICONOCLAST[4] (Power, 2000; Bouayad-Agha *et al.*, 2000, 2001), la *Document Structure* est implémentée sous la forme d'un arbre ordonné où chaque nœud a une catégorie textuelle. De manière analogue à la grammaire de Nunberg, une hiérarchie de catégories est utilisée. Les auteurs suggèrent que cette hiérarchie peut différer d'un type de document à l'autre, mais qu'elle est toujours présente. Dans ICONOCLAST, cinq catégories sont utilisées, allant du niveau 0 (le plus bas dans la structure) au niveau 5 (le plus élevé) :

0 *text-phrase*
1 *text-clause*
2 *text-sentence*
3 *paragraph*
4 *section*
5 *chapter*

[2] Nunberg parle de « "null" elements like spacing to separate text elements like words and lines from one other » (Nunberg, 1990, p. 52).
[3] Phrases syntaxique et visuelle peuvent correspondre, mais cela n'est pas systématique.
[4] http://www.nlg-wiki.org/systems/ICONOCLAST

Une *text-phrase* est une expression qui est davantage contrainte par la grammaire traditionnelle que par la grammaire visuelle (Nunberg, 1990, p. 33). Cette catégorie se retrouve notamment dans les expressions entre parenthèses ou tirets cadratins.

Les auteurs proposent de généraliser les règles de réécriture de Nunberg en introduisant le symbole L_N qui représente la catégorie L associée au niveau N :

$$L_N \rightarrow L^+_{N-1} \quad (N > 0)$$

Ainsi l'unité d'une catégorie donnée est uniquement constituée d'unités de catégories directement inférieures, et seules des unités de catégories identiques peuvent être coordonnées.

Pour pouvoir modéliser des documents plus complexes, les auteurs introduisent également un mécanisme d'indentation. L'intégration de l'indentation à la syntaxe dénote la volonté de considérer que l'indentation relève de la structure logique (ici la *Document Structure*) et non de la représentation physique. La justification en est qu'un élément peut être indenté à un autre sans qu'il y ait pour autant une indentation graphique[5]. Les auteurs utilisent le terme d'indentation logique (*logical indentation*) (Power *et al.*, 2003, p. 225 - 226).

Dans la pratique, l'indentation permet de représenter des apparentes violations de la hiérarchie telles que, par exemple, le fait qu'une phrase textuelle subordonne un paragraphe[6]. Formellement, les valeurs de l'indentation varient entre $I_0 \ldots I_{MAX}$ où I_0 est l'unité sans indentation et I_{MAX} est l'unité la plus imbriquée dans un modèle de document donné. Cette notation permet de réécrire la généralisation pour les éléments non-indentés :

$$[L_N, I_M] \rightarrow [L_{N-1}, I_M]^+$$

où $[L_N, I_M]$ une catégorie de niveau N et d'indentation M. Pour les éléments indentés, il est alors possible d'introduire la règle de réécriture suivante :

$$[L_A, I_M] \rightarrow [L_B, I_{M+1}]^+$$

où L_B peut représenter une catégorie de niveau plus élevé que L_A.

Nous donnons dans l'exemple (2.a) et la figure 2.1 deux exemples issus de Power *et al.* (2003). L'exemple (2.a) est une structure énumérative (qui sera également étudiée dans le chapitre 3). La figure 2.1 représente la structure logique correspondant à cet exemple selon le modèle discuté ici. Les nœuds de l'arbre sont étiquetés du nom des catégories et de leur niveau d'indentation associé (entre parenthèses).

[5] Une indentation pourrait être obtenue en utilisant des contrastes de fontes (p. ex. mise en italique).
[6] Ce type de construction est notamment courant dans les notices de médicament.

(2.a)

In rare cases the treatment can be prolonged for another week ; however, this is risky since

- The side-effects are likely to get worse. Some patients have reported severe headache and nausea.
- Permanent damage to the liver might result.

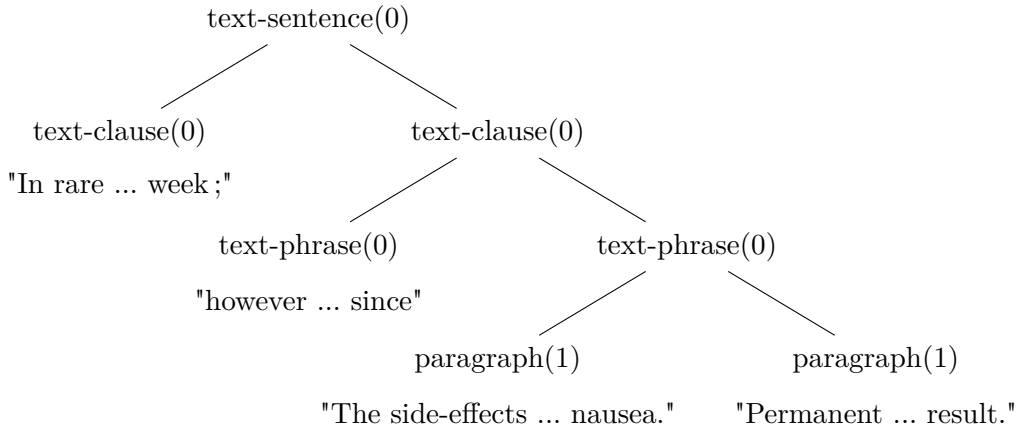

FIGURE 2.1 : Exemple d'arbre représentant la structure de document selon Power *et al.* (2003) pour l'exemple (2.a)

Lien avec la structure discursive Dans le cadre de la génération de textes, le lien entre la structure du document et la structure discursive a été initialement étudié par Scott et de Souza (1990). Les auteurs reprennent cette tâche : il s'agit de transformer un *message*, exprimé sous la forme d'un arbre rhétorique non-ordonné, en un texte structuré dont les éléments sont ordonnés par l'ordre de lecture. La difficulté réside dans le fait que les structures rhétoriques et logiques ne sont pas isomorphiques[7]. Une même structure rhétorique peut être réalisée par de multiples structures logiques.

Power (2000) propose qu'une génération textuelle soit adéquate lorsqu'elle rencontre trois conditions : (i) un contenu correct (toutes les propositions de la structure rhétorique sont exprimées), (ii) une structure correcte au vu de la syntaxe (présentée précédemment), et (iii) une compatibilité structurale (*structural compatibility*).

Sur base de ce dernier critère, ne sont *compatibles* que les structures de document où les regroupements de propositions sont également présents dans la structure rhétorique. Bouayad *et al.* (2000) reformulent :

[7] Selon Bouayad *et al.* (2000) il y a isomorphisme lorsque « (...) every set of propositions that is dominated by a node in DocRep should be dominated by a node in RhetRep, and vice-versa ». Voir également (Power *et al.*, 2003, p. 235).

> Formally, every set of propositions that is dominated by a node in DocRep should be dominated by a node in RhetRep — but the converse is not required.

où DocRep et RhetRep désignent respectivement la structure de document et la structure rhétorique. Cette compatibilité structurale permet de remplacer le critère d'isomorphie considéré comme trop rigide.

Dans la pratique, le système ICONOCLAST prend en entrée un arbre rhétorique[8] exprimé selon le formalisme de la *Rhetorical Structure Theory* (RST)[9]. Ensuite, la génération des structures de document est considérée comme un problème de satisfaction de contraintes (CSP). Les contraintes traduisent les trois conditions de Power, ainsi que l'ordre des connecteurs de discours utilisés. Finalement, la qualité des structures générées est évaluée selon des propriétés stylistiques (p. ex. l'ordre des constituants en fonction de la relation portée).

Ainsi, la forme linguistique finale d'un texte généré est réalisée à la fois par la structure de document et par des propriétés syntaxiques qui interviennent essentiellement au niveau de la clause.

2.1.2 Modèle de Bateman *et al.* (2001)

Le modèle de Bateman *et al.* (2001) propose également une structure logique abstraite. Cette structure joue un rôle central dans DArt_{bio}[10], pour *Dictionary of Art: biographies*, un système de génération de biographies. Bateman *et al.* (2001) envisagent les biographies comme des documents multimodaux, où les éléments graphiques et textuels entretiennent des relations complexes. Leur modèle propose alors de représenter simultanément ces deux types d'éléments.

Dans la suite de la section, nous présentons d'abord le modèle théorique avant d'en décrire la mise en œuvre dans la procédure de génération des biographies.

Modèle théorique Trois structures sont considérées dans le modèle de Bateman *et al.* (2001) : la structure physique, la structure logique et la structure rhétorique. Nous présentons ces trois structures ci-dessous.

[8] Notons que dans le travail Bouayad-Agha *et al.* (2001) une méthode utilisant une représentation rhétorique non-hiérarchique est proposée.

[9] La *Rhetorical Structure Theory* (RST) de Mann et Thompson (1988) analyse l'organisation des textes du point de vue de leur cohérence. Cette organisation est reflétée par un arbre de constituants : les nœuds terminaux représentent les segments textuels et les nœuds non-terminaux portent le nom des relations rhétoriques liant leurs constituants. Il existe deux familles de relations rhétoriques : (i) les relations multi-nucléaires qui lient des constituants d'importance égale et (ii) les relations noyau-satellite qui lient un noyau dont le contenu propositionnel est saillant à son satellite. Le choix des relations se fait sur la base d'un critère fonctionnel (et non lexical ou syntaxique).

[10] http://www.nlg-wiki.org/systems/Dart-bio

La structure physique (*page layout*) est composée de blocs visuels (*visual blocks* ou *layout forms*) dont les auteurs ne donnent pas explicitement une définition[11]. Les auteurs laissent cependant entendre que ces blocs recouvrent des éléments textuels (p. ex. paragraphes, listes à puces, etc.)[12] ou graphiques (p. ex. diagrammes, figures, etc.). Les éléments de plus bas niveau tels que la ponctuation, la fonte ou les espacements sont assimilés à de la microtypographie et ne sont pas pris en compte.

La structure abstraite de document (*layout structure*) organise les blocs visuels selon un critère de regroupement visuel. Bateman *et al* (2001) expliquent :

> Layout structure abstracts across the precise details of physical layouts to focus on classes of layouts that are visually "equivalent." Visually equivalent layouts suggest the same page blocks, with similar inter-block relationships of perceived prominence and similarity.

Au sein de cette structure, les blocs visuels prennent le nom d'unités de mise en forme (*layout unit*) et sont liées par des relations typographiques telles que l'emboîtement (*containment*)[13], l'ordre de lecture (*reading order*), la similarité (*similarity*) ou la référence (*reference*)[14]. La relation d'emboîtement permet de créer l'ossature de l'arbre de constituants représentant la structure de document. Les nœuds non-terminaux sont des conteneurs abstraits, regroupant des parties plus ou moins vastes du document, étiquetés ou non, tandis que les nœuds terminaux correspondent aux blocs visuels.

Dans la figure 2.2, nous donnons un exemple de correspondance entre une page de magazine et sa structure de document. Cet exemple est issu du travail de Reichenberger *et al.* (1996), repris par Bateman *et al.* (2001). Les étiquettes utilisées pour les nœuds non-terminaux sont définies de manière *ad hoc* par rapport au sujet dont traite le document. Ici il s'agit de la description d'un sport (le floorball[15]) et les étiquettes sont définies en conséquence (*rules*, *equipment*, etc). Les nœuds terminaux peuvent à la fois être des éléments textuels (p. ex. `var.2.2.2`) et des éléments graphiques (p. ex. `eq.1`). Les liens hiérarchiques correspondent à la relation d'emboîtement. Les liens horizontaux correspondent aux relations de référence, d'ordre de lecture et de similarité.

La structure discursive (*rhetorical structure*) modélise les intentions communicationnelles. L'approche multimodale du travail rend difficile la définition d'un ordre de lecture pour les documents complexes (p. ex. figure 2.2). Les auteurs ont donc été amenés à utiliser la RST. Celle-ci est indépendante de l'ordre de lecture et permet une représentation en arbre adaptée aux mises en forme complexes (p. ex. multiples colonnes, images, etc.).

[11] « (...) we are most concerned with macrotypography–the segmentation of a page of information into more or less closely related "visual blocks." » (Bateman *et al.*, 2001)

[12] « Examples of textual layout forms are textblocks consisting of paragraphs, enumerated lists, itemized lists, and the like. » (Bateman *et al.*, 2001)

[13] Voir également le travail de Southall (1989) pour la relation d'emboîtement (*containment*).

[14] Lorsque deux blocs visuels sont liés par une proximité physique sur la page.

[15] Il s'agit essentiellement d'une variante du hockey sur glace dans laquelle le terrain est en linoleum ou en polymère et le palet est remplacé par une balle.

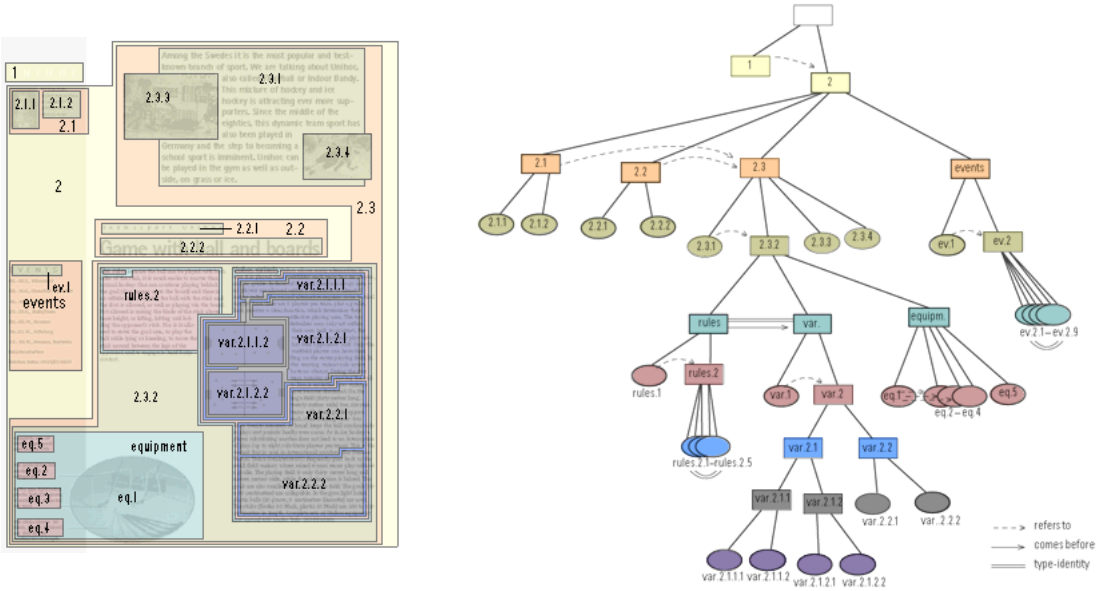

FIGURE 2.2 : Exemple de page de magazine segmentée en blocs visuels et sa structure logique selon Reichenberger *et al.* (1996) et Bateman *et al.* (2001)

Une correspondance est recherchée entre les sous-arbres de la structure logique et les sous-arbres de la RST. Néanmoins, la décomposition de l'arbre RST n'est pas immédiate. De manière analogue à Power *et al.* (2003), les auteurs soulignent que l'isomorphie entre les deux arbres n'est pas automatique :

> Mapping is generally achieved by placing parts of the RST-structure in correspondence with particular nodes in a layout structure. This proceeds recursively down through the RST tree. As we have now seen, however, the correspondence is complicated by the fact the layout structure and the RST tree need not remain congruent.

Ainsi, la structure rhétorique est également vue ici comme distincte de la structure abstraite de document et les liens entre les blocs visuels ne trouvent pas une correspondance immédiate en termes discursifs[16].

Génération des textes Le système DArt$_{bio}$ utilise les informations de l'encyclopédie *Dictionary of the Art* (Turner, 1996) formatées selon le modèle de domaine de Rostek *et al.* (1994). Celles-ci sont ensuite fournies à plusieurs modules : le générateur de diagrammes AVE (Reichenberger *et al.*, 1995), le générateur de langage KOMET[17] (Bateman et Teich, 1995) et, enfin, le générateur de mise en forme APALO.

[16] « A page's most salient features are visual –i.e., typographical– and part of our claim is that this is not directly indicative of rhetorical organization : relationships between visual blocks on the page are at a different level of abstraction than rhetorical relations. » (Bateman *et al.*, 2001)

[17] http://www.nlg-wiki.org/systems/KOMET

Le module KOMET utilise des structures préalablement construites et adaptées au genre biographique pour générer un arbre RST. Les nœuds sont complétés par des informations (p. ex. naissance, carrière, etc.). À partir de cet arbre RST, une procédure de descente récursive est entamée par le module APALO afin de construire l'arbre de la structure de document. À chaque nœud de l'arbre RST, le module cherche une correspondance entre le sous-arbre RST et une unité de mise en page. La présence d'une correspondance est déterminée ou infirmée sur base d'heuristiques et de contraintes. Une difficulté réside dans la condition d'arrêt de la procédure. Les auteurs soulignent la nécessité d'un compromis entre textuel et visuel[18].

Les résultats obtenus sur cinq biographies montrent que la transposition est très largement non déterministe. Les auteurs suggèrent la nécessité d'adjoindre des contraintes empiriquement motivées. Le projet GeM[19] (pour *Genre and Multimodality*) s'inscrira dans cette voie en proposant un schéma d'annotation complexe (5 niveaux de structures, 3 types de contraintes) pour le traitement des documents multimodaux (Allen *et al.*, 1999; Bateman et Delin, 2001; Delin *et al.*, 2002).

2.1.3 Modèle de Virbel (1989)

Le modèle de Virbel (1989), aussi appelé Modèle d'Architecture Textuelle (MAT), présente une structure abstraite des textes, l'*Architecture de texte*, construite à l'aide d'un métalangage formalisé de manière semblable à celui de Harris (1968). Les travaux de Pascual (1991) et de Luc (2001) ont proposé par la suite des extensions à ce modèle.

Une notion introduite ici concerne le continuum existant entre une formulation discursive et une formulation qui utilise la mise en forme. Virbel (1989) explique :

> For example, whether I write *I shall introduce my subject by saying A* or *INTRODUCTION : A*, I am asking my reader to treat textual segment *A* as an *introduction* to what follows.

Dans la suite de cette section, nous introduisons le métalangage de Harris, ensuite nous présentons le MAT en lui-même et, enfin, nous montrons son lien avec la structure discursive.

Métalangage Pour Harris (1968), le métalangage est l'ensemble des métaphrases (*metalinguistic sentence*), c'est-à-dire des phrases qui ont pour sujet le langage ou une production de celui-ci. Harris explicite et justifie :

> Every natural language must contain its own metalanguage, i.e., the set of sentences which talk about any part of the language, including the whole grammar of the language. Otherwise, one could not speak in a language about that language itself; this would conflict with the observation that in any language one can speak about any subject, including the language and its sentences, provided that required terms are added to the vocabulary.

[18] « (...) the more the RST structure is decomposed, resulting in typographical distinctions, the less use is made of explicit linguistic discourse marking. » (Bateman *et al.*, 2001)

[19] http://www.purl.org/net/gem

Ainsi, la phrase « Le chat dort » peut être décrite par la métaphrase « 'Le chat dort' est une phrase ». Il peut exister également des métaphrases qui ne citent pas de segment telle que « Les phrases en français contiennent un verbe » [20]. Ces métaphrases peuvent apparaître dans des phrases naturelles (comme nous le faisons dans ce paragraphe) ou bien également être l'argument de métaphrases de niveau supérieur tel que par exemple « 'Le chat dort' est une phrase est une phrase ».

Le métalangage est toutefois limité par son incapacité à définir des contraintes d'une langue, car lui-même décrit par cette langue (Harris, 2002). Intervient ici la notion de système transformationnel[21]. Comme toute autre phrase de la langue, les métaphrases linguistiques peuvent être réduites en phrases élémentaires (*elementary sentences*)[22], ou, inversement, peuvent être dérivées par l'application d'opérateurs de base (*base operators*)[23] (Luc, 2000). Ces transformations laissent des traces dans la phrase d'arrivée. Par exemple, « Veux-tu du café ? » est le résultat de la réduction de « Je te demande si tu veux du café » et où les traces laissées sont le point d'interrogation et l'intonation[24]. Notons que l'absence de distinction entre oral et écrit est un héritage de Harris.

Ainsi, les phrases élémentaires et les opérateurs permettent l'application de contraintes sur le langage et sont, par leur fonction, porteurs d'une interprétation linguistique[25].

Le modèle théorique Le MAT définit deux structures qui font chacune appel à une notion : la réalisation physique est décrite par la *Mise en Forme Matérielle* et la structure abstraite de document est formalisée au sein de l'*Architecture de texte*.

La *Mise en Forme Matérielle* (MFM) (Virbel, 1985) permet la description du niveau visuel des textes en considérant des indices de types typographiques et dispositionnels, lexicaux et syntaxiques. Pour les deux premiers types d'indices, la MFM se rapproche de la grammaire de Nunberg. La MFM cherche à théoriser les éléments de mise en page de relativement bas niveau (p. ex. fontes, marges, etc.), et, dans les deux cas, l'intérêt se porte davantage sur l'analyse du langage plutôt que sa génération. Néanmoins, la MFM s'écarte du travail de Nunberg en considérant également les indices lexicaux et syntaxiques. Ce choix introduit la notion d'équivalence entre des éléments purement discursifs et des éléments typographiques et dispositionnels.

[20] Exemples adaptés de (Harris, 1968, p. 127).

[21] Chomsky s'inspirera du système transformationnel de Harris pour mieux le réinventer dans son propre travail (Chomsky, 1957).

[22] Pour Harris, une phrase élémentaire a la forme $N\ V\ \Omega$ où N est un nom, V un verbe et Ω ses compléments (Harris, 1968, p. 68).

[23] Un opérateur de base permet d'appliquer une transformation élémentaire entre deux phrases. Par exemple, l'opérateur ϕ_p permet la permutation (Harris, 1968, p. 77).

[24] Exemple adapté conjointement de (Virbel, 1989) et de (Luc, 2000).

[25] « Furthermore, the primitive eIements —elementary sentences K and operators ϕ— have meanings, i.e., a linguistic interpretation, such that the meaning of a sentence, as a sequence of K and ϕ, is the sequence of meanings of its component K and ϕ. No major independent semantic theory is thus needed. » (Harris, 1968, p. 2).

Ceci amène une problématique où (i) plusieurs formulations physiques peuvent représenter un même objet linguistique et, inversement, (ii) une même mise en forme peut endosser des rôles différents. L'exemple largement repris par les auteurs est celui de la définition dont l'expression suit un continuum entre le discursif et le typo-dispositionnel (Virbel, 1989; Pascual et Péry-Woodley, 1995; Luc, 2000).

L'*Architecture de texte* constitue une réponse face à la variabilité des indices visuels en considérant uniquement les unités de mise en forme au travers de leur formulation discursive. Pascual et Virbel (1996) explicitent :

> In fact, the combination of all possible typographic, positional and syntactic variations for a textual object (like a definition) leads to a number of possible forms of the order of one million. (...) instead of studying millions forms of the same textual object, the study is restricted to only running text, expressing explicitly the textual object corresponding to the formulation. In this case linguistic methods of production and analysis can be used.

Le versant discursif est représenté au travers d'un métalangage similaire à celui de Harris et introduisant deux unités :

- l'**objet textuel** est « un segment caractéristique de texte, rendu perceptible par un jeu de contrastes de la mise en forme matérielle » (Pascual et Péry-Woodley, 1995). Ce segment peut être une section, un paragraphe, une énumération, une définition, etc.
- l'**unité textuelle** est « un segment de texte ne comportant aucun objet textuel, c'est-à-dire un segment de texte entièrement discursif. » (Luc, 2000). Cette unité est l'unité la plus petite du métalangage.

Ces deux unités sont organisées au travers de métaphrases Harrissiennes. Dans ce cadre, elles sont la forme discursive d'un phénomène de mise en forme. Leur construction se fait sur la base de verbes performatifs dont l'acte illocutoire (au sens d'Austin (1975) et Searle (1976)) est tourné vers le texte. Plusieurs types de métaphrases ont été considérés par les auteurs avec des rôles divers (p. ex. organisation du texte, inclusion d'objets textuels, mise en relation d'unité textuelle et d'objet textuel, etc.). L'ensemble des métaphrases représentant un document est appelé un métadiscours.

Nous présentons ici un exemple sous la forme d'une image de texte[26] (Figure 2.3)[27]. À gauche, les objets textuels sont identifiés par leur type et un numéro d'identifiant unique au sein de chaque type. À droite, les unités textuelles (UT) sont suivies d'un identifiant unique. L'architecture de cet exemple est représenté par le métadiscours donné dans la figure 2.4.

[26] L'image de texte est une notation introduite par Pascual (1991).
[27] Cet exemple est adapté de Luc (2000).

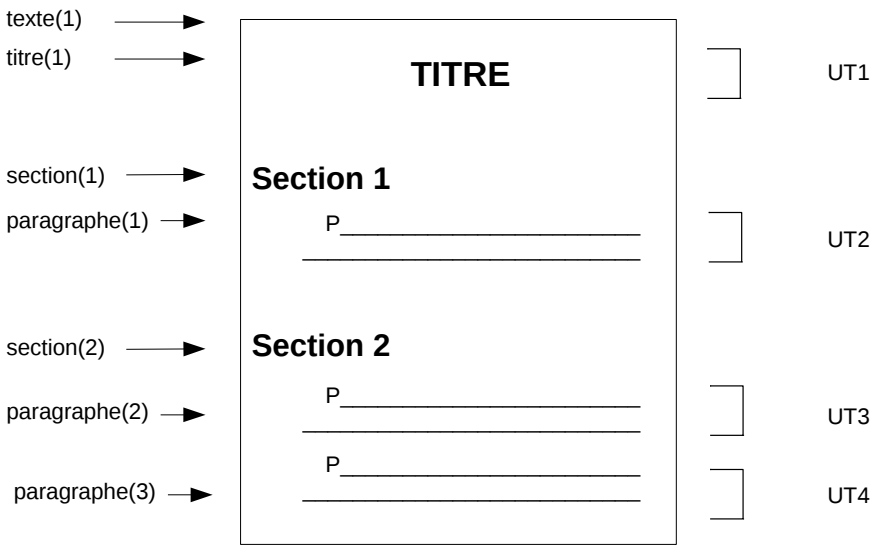

FIGURE 2.3 : Exemple d'image de texte

```
L'auteur crée un texte identifié texte(1).
    L'auteur intitule texte(1) par un titre identifié titre(1).
    L'auteur attache unité_textuelle(1) à titre(1).
    L'auteur compose titre(1) de unité_textuelle(1).
    L'auteur organise texte(1) en 2 parties identifiées section(1), section(2).
    L'auteur numérote section(1), section(2).
        L'auteur développe un paragraphe identifié paragraphe(1).
        L'auteur attache unité_textuelle(2) à paragraphe(1).
        L'auteur compose paragraphe(1) de unité_textuelle(2).
    L'auteur compose section(1) de paragraphe(1).
        L'auteur développe un paragraphe identifié paragraphe(2).
        L'auteur attache unité_textuelle(3) à paragraphe(2).
        L'auteur compose paragraphe(2) de unité_textuelle(3).
        L'auteur développe un paragraphe identifié paragraphe(3).
        L'auteur attache unité_textuelle(4) à paragraphe(3).
        L'auteur compose paragraphe(3) de unité_textuelle(4).
    L'ateur compose section(2) de paragraphe(2), paragraphe(3).
L'auteur compose texte(1) de titre(1), section(1), section(2).
```

FIGURE 2.4 : Métadiscours de l'image de texte en figure 2.3

Le métadiscours peut également être représenté sous la forme d'un *graphe architectural*. Cette représentation en graphe permet de traiter les phénomènes qui ne sont pas purement hiérarchiques[28]. Au sein de ce graphe, l'arbre formé par la relation de composition (verbe performatif *composer*) est un arbre de constituants où les nœuds non-terminaux sont les objets textuels et les nœuds terminaux sont les unités textuelles (Figure 2.5).

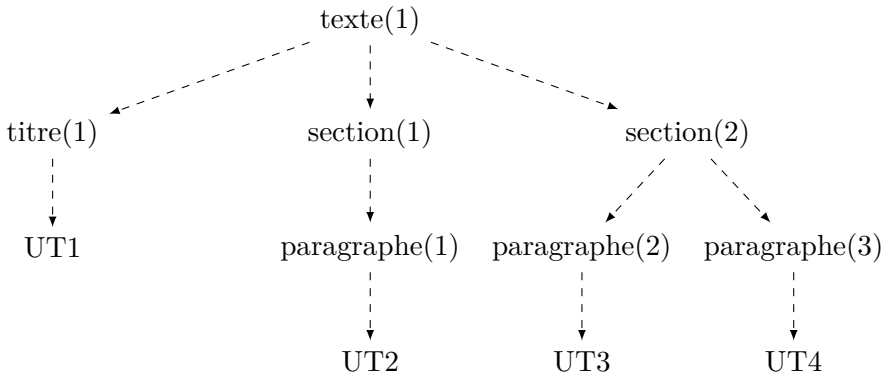

FIGURE 2.5 : Graphe architectural correspondant à l'image de texte en figure 2.3 et le métadiscours en figure 2.4

La cohérence du métadiscours est assurée par des contraintes portant sur la relation de précédence (p. ex. une unité ne peut en composer une autre que si elle a précédemment été introduite) ou la relation de composition. Pour cette dernière, les objets textuels suivent un ordre d'inclusion de manière analogue à la hiérarchie du modèle de Power *et al.* (2003). Ici, cet ordre est exprimé au travers de larges tables de composition, initialement proposées par Pascual (1991). Elles seront reprises et étendues par Luc (1998).

Le recours à un métalangage permet de s'appuyer sur une approche linguistique de la structure (tant au niveau de la génération que de l'analyse). Les métaphrases utilisées mettent en relation des actes de langage (tournés vers le texte), des objets textuels et des unités textuelles. Dans ce contexte, les marques typo-dispositionnelles peuvent alors être considérées comme des traces de réduction d'une forme purement discursive du texte (dit *prototexte* (Virbel, 1989)).

Lien avec la structure discursive Le lien entre l'Architecture de texte et la structure discursive, exprimée au travers de la RST, a été étudié par Luc (2000). Celui-ci propose de composer les deux modèles afin de bénéficier des apports de chacun.

D'un côté, l'Architecture de texte permet aisément de représenter des objets textuels, mais il est difficile de représenter les relations qui lient ces objets. D'un autre côté, la RST permet une représentation hiérarchique des objets, mais celle-ci est uniquement guidée par des critères fonctionnels (indépendamment de l'ordre de lecture) et n'est pas adaptée aux phénomènes non-hiérarchiques (p. ex. structures entrelacées).

[28] Cette notion sera à nouveau discutée dans le chapitre 4.

Luc propose de représenter conjointement la RST et l'Architecture de texte en logique des prédicats. Les prédicats pour la RST organisent les segments textuels, tandis que ceux de l'Architecture de texte organisent les objets textuels.

Parmi les objets textuels, une distinction est faite entre les *objets textuels fonctionnels*, qui sont des objets textuels participant à la structure discursive du texte généralement sous la forme de segments minimaux (p. ex. items de structures énumératives, titres, amorces, exemples, etc.), et les *objets textuels structurels*, qui peuvent participer au discursif mais pas de manière systématique (p. ex. paragraphe, section, etc.). Dans ce cadre, l'auteur propose de faire un lien uniquement entre les segments de texte de la RST et les objets textuels fonctionnels de l'Architecture de texte.

Ainsi, au sein de la RST, certains segments ont le statut d'objet textuel et au sein de l'Architecture de texte, certains objets textuels sont liés par des relations rhétoriques. Il est alors possible d'extraire deux sous-graphes partiels représentant respectivement la structure rhétorique du texte et l'Architecture de texte (Luc, 2000, p. 149). Luc exploite cette composition entre modèles au travers d'un système de génération des textes.

2.1.4 Comparaison entre les modèles théoriques

Cette section propose une comparaison entre le modèle de Power *et al.* (2003), ci-après MDS, le modèle de Bateman *et al.* (2001), ci-après MBAT, et le modèle de Virbel (1989), ci-après MAT. Trois points sont considérés :

- leur objectif commun en génération de textes ;
- le traitement de la relation d'inclusion ;
- la granularité des phénomènes étudiés.

Objectif commun en génération de textes Les trois modèles présentés interviennent essentiellement dans le cadre de la génération de textes et trouvent leur origine dans le travail pionnier de Hovy et Arens (1991) qui associe le formatage visuel à la génération. Cette association est néanmoins faite sans la médiation d'un modèle abstrait de document. En effet, Hovy et Arens font une correspondance directe entre des séquences de relations rhétoriques et des commandes LaTeX. Les auteurs expliquent :

> An additional shortcoming with our approach is the fact that we embed LaTeX commands literally into the RST plans. The text structure planner thus has no ability to reason about the implications of its formatting. Better would be to develop an abstract representation of textual devices which, when included in the text plans, would be realized into LaTeX (...) commands at the time the content is realized into English.

Dans la suite, le MDS, le MBAT et le MAT adhèrent à cette nécessité d'un niveau abstrait situé entre le physique et le rhétorique. Leurs terminologies respectives reprendront ce choix (Table 2.1) et une problématique identique sera abordée : un même message (exprimé par la RST) peut être *matérialisé*[29] par différents dispositifs de mise en forme,

[29] Ce terme est repris de la terminologie du MAT. Power *et al.* (2003) utilisent le verbe *to realize*.

et le choix d'un dispositif implique une adaptation de l'aspect syntaxique. Par exemple, une énumération horizontale présente généralement davantage d'éléments syntaxiques et moins de ponctuation qu'une énumération verticale.

structures	MDS Power *et al.* (2003)	MBAT Bateman *et al.* (2001)	MAT Virbel (1989)
visuelle	Physical representation	Page layout	Structure visuelle
logique	Document structure	Layout structure	Architecture de texte
discursive	Rhetorical structure	Rhetorical structure	Structure discursive

TABLE 2.1 : Comparaison des terminologies utilisées pour la désignation des différentes structures au sein des modèles théoriques de structuration de document

Ce cadre en génération de textes implique qu'un intérêt commun est porté à la structure abstraite et à son lien avec la structure discursive. Dans la pratique, cela se traduit par un travail fin (généralement manuel ou semi-automatique) et sur des genres restreints de textes. Pour le MDS, Power *et al.* (2003) ont travaillé sur des posologies de médicaments. Dans le cadre du MBAT, Bateman *et al.* (2001) analysent manuellement des pages d'un magazine de sport. Dans le MAT, Luc (2000) propose une analyse d'un texte à consignes. Dans les trois cas, peu de résultats empiriques sont donnés, et les contributions sont essentiellement théoriques.

Ces choix s'opposent aux approches dans la communauté d'Analyse du Document où les contributions portent sur le passage automatique entre la structure visuelle et la structure logique. Ces approches sont discutées dans la section suivante (section 2.2).

Relation d'inclusion La relation d'inclusion entre les unités de la structure logique est essentielle au sein des modèles. Néanmoins, la manière de l'exprimer et les contraintes qui lui sont associées ont des implications quant au formalisme de représentation.

- Dans le MDS, l'inclusion est exprimée au travers de règles de réécriture. Celles-ci portent sur des catégories textuelles (p. ex. clause textuelle, phrase textuelle, paragraphe, etc.) et permettent l'expression d'une syntaxe stricte : (i) une catégorie textuelle donnée est uniquement constituée de catégories qui lui sont directement inférieures[30], et (ii) seules des catégories identiques peuvent être coordonnées. Dans ce contexte, le MDS est représenté au travers d'un arbre de constituants où les nœuds non-terminaux sont les catégories textuelles endossant le rôle de *tête* et les nœuds terminaux sont les *constituants* auxquels sont associés un contenu textuel.

- Dans le MBAT, l'inclusion est exprimée par la règle typographique d'emboîtement (*containment*). Celle-ci lie des unités de mise en forme (*layout unit*) à des blocs visuels selon un critère spatial. D'autres relations sont également considérées : similarité, ordre de lecture, etc. La représentation de la structure abstraite est faite

[30] Notons que l'ajout de l'indentation permet néanmoins de contrevenir à cette règle pour la description de phénomènes tels que les citations, les listes à puces, etc. Voir section 2.1.1.

au travers d'un arbre de constituants[31]. Les nœuds non-terminaux sont des containeurs abstraits étiquetés (p. ex. *rules*) ou non (p. ex. coin en haut à droite de la page) et les nœuds terminaux sont les blocs visuels.

- Dans le MAT, l'inclusion est représentée par la métaphrase du verbe *composer*. Celle-ci porte sur deux types d'unités : les objets textuels (p. ex. paragraphe, énumération, théorème, etc.) et les unités textuelles (segments textuels non mis en forme). L'ordre hiérarchique est établi au travers de tables de composition. La relation de composition intervient au sein d'un ensemble de métaphrases appelé métadiscours. Ce métadiscours peut être traduit sous la forme d'un graphe (dit *architectural*) au sein duquel l'arbre formé par la relation de composition est un arbre de constituants. Ses nœuds non-terminaux sont des objets textuels et ses nœuds terminaux sont des unités textuelles. La propriété de graphe permet au MAT de traiter des phénomènes non hiérarchiques (p. ex. énumérations entrelacées).

La table 2.2 reprend les différentes représentations utilisées pour représenter la structure de document, ainsi que celles utilisées pour les structures visuelles et rhétoriques.

structures	MDS Power *et al.* (2003)	MBAT Bateman *et al.* (2001)	MAT Virbel (1989)
visuelle	Grammaire de Nunberg	-	Mise en Forme Matérielle
logique	Arbre de constituants	Arbre de constituants	Graphe architectural
discursive	Arbre RST	Arbre RST	Arbre RST

TABLE 2.2 : Comparaison des représentations utilisées pour les différentes structures au sein des modèles théoriques de structuration de document

Granularité des phénomènes étudiés Il est difficile d'établir des frontières nettes quant à la granularité des phénomènes étudiés par les modèles. Néanmoins, nous pouvons avancer que, généralement, le MDS porte son intérêt à un niveau davantage intra-phrastique, tandis que le MBAT et le MAT s'intéressent à un niveau ultra-paragraphique, voire ultra-textuel pour le second. Nous schématisons (et nécessairement simplifions) le propos en positionnant les modèles le long d'un continuum entre la clause et le texte dans la figure 2.6 [32]. Nous donnons ci-dessous deux commentaires respectivement à propos de la frontière haute et de la frontière basse de la figure.

- Au niveau supérieur, le MBAT considère le texte dans son intégralité matérielle (p. ex. figures, numéro de page, en-têtes, etc.). Ceci s'explique par l'intérêt que Bateman *et al.* (2001) portent aux document multimodaux, où éléments textuels

[31] À proprement parler, il s'agit d'un graphe : « We represent layout structures in terms of a tree structure (representing containment) augmented by a restricted set of possible additional annotations corresponding to the remaining typographical relation types. » (Bateman *et al.*, 2001)

[32] Notons que l'alinéa est défini ici comme un segment textuel encadré par deux moyens dispositionnels, mais permettant néanmoins une interprétation linguistique (p. ex. nom d'auteur, etc.).

et graphiques sont en relation. Ce choix du genre de document (*document genre*)[33] implique un grain plus grand à trois niveaux : (i) les unités sont des blocs visuels généralement non étiquetés, (ii) les étiquettes proposées ne sont pas génériques et (iii) la représentation formelle de la structure est non contrainte (p. ex. absence de hiérarchie, de règles, etc.). Cette position contraste avec celles prises par le MDS et le MAT au sein desquelles le *corps de texte* est central et sa représentation contrainte.

- Au niveau inférieur, la prise en compte par le MDS d'indicateurs de bas niveau (p. ex. ponctuation, fontes, espaces, etc.) lui permet de manipuler des unités très fines. Par exemple, Power *et al.* (2003) distinguent une unité au sein de la clause[34]. Cette décomposition ouvre la voie à la caractérisation de schémas englobant plusieurs unités où l'indentation intervient (p. ex. citations, énumérations). Le MAT se situe légèrement plus haut en considérant le segment textuel comme unité atomique. Dans son travail, Luc propose une étude de la composition de ces segments en structures complexes (imbriquées, entrelacées, etc.). *A contrario*, Bateman *et al.* (2001) n'étudient que très peu la composition des unités complexes[35].

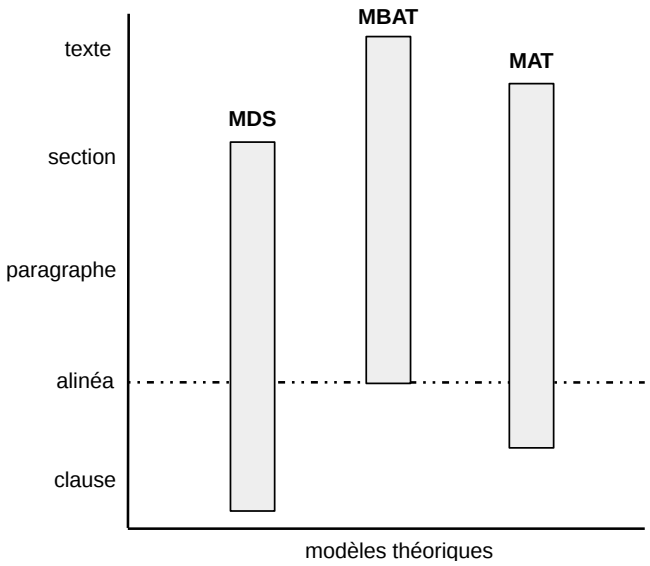

FIGURE 2.6 : Schématisation de la granularité des phénomènes étudiés au sein des modèles de structuration de document

[33] Ce terme est introduit dans le projet GeM (Allen *et al.*, 1999; Bateman et Delin, 2001).
[34] Cette unité est la *text-phrase*, qui est généralement marquée par des parenthèses ou des tirets.
[35] « Our heuristics do not bother to decompose when there is little content to be expressed ; and as each of the satellites in the biography contains only a few propositions, we do not decompose further here, selecting a single layout unit with a textblock layout form. » (Bateman *et al.*, 2001)

2.2 Approches empiriques en Analyse du Document

La dichotomie entre structure physique et structure logique est traditionnellement faite dans la description des documents depuis les années 80 (Furuta *et al.*, 1982; André *et al.*, 1989, 1990). Cette vision a notamment été définie dans le standard international ISO 8613 (1989). Dans ce paradigme, le rôle de l'auteur est considéré comme limité à la structure logique et la mise en page est laissée au soin d'un éditeur, d'un metteur en page ou d'un module informatique. Furuta (1989) explicite :

> (...) the specification of the logical structure is generally provided by the document's author and the specification of the layout is generated by the formatting process.

Cette distinction est originellement reprise par la communauté d'*Analyse du Document* (conférences ICDAR[36], CIFED[37], etc.) où chacune de ces structures appelle une tâche spécifique (Mao *et al.*, 2003) :

- la structure visuelle, aussi dite structure géométrique, résulte de l'analyse géométrique (*document analysis*).
- la structure logique résulte de l'analyse logique (*document understanding*). Cette tâche se situe généralement après l'analyse visuelle.

2.2.1 Analyse géométrique

L'objectif de l'analyse géométrique est de segmenter le document en zones visuellement homogènes (Paaß et Konya, 2012). Cette tâche suit la définition de la structure géométrique donnée par le standard ISO 8613 (1989) :

> The result of dividing and subdividing the content of a document into increasingly smaller parts, on the basis of the presentation, for example, into pages, blocks.

Les étiquettes de ces parties de documents peuvent être définies (p. ex. texte, image, tableau, etc.), néanmoins celles-ci restent indissociables de la tâche donnée et du domaine dans lesquels celles-ci s'inscrivent[38].

Dans cette section, nous reprenons la classification faite par Paaß et Konya (2012) en divisant l'analyse géométrique selon trois familles d'approches : les approches descendantes, les approches ascendantes et, enfin, les approches hybrides.

[36] International Conference on Document Analysis and Recognition
[37] Colloque International Francophone sur l'Écrit et le Document
[38] « It is important to note that in the specialized literature there exists no consensus on the number of physical classes considered, the number depends mostly on the target domain. » (Paaß et Konya, 2012)

Approches descendantes Les approches descendantes consistent à diviser récursivement le document jusqu'à obtenir des unités « atomiques »[39] ou jusqu'à rencontrer une condition d'arrêt préalablement définie. Ces approches sont généralement rapides et efficaces sous la condition que certaines informations aient été fournies préalablement[40]. Ces informations concernent la nature des documents (p. ex. lettres, dossiers de patients, etc.) ou leur format. Les méthodes généralement utilisées reposent sur la recherche de plages blanches (Pavlidis et Zhou, 1991), les profils de projection (Ha *et al.*, 1995a) ou encore l'algorithme RXYC (*Recursive X-Y Cuts*) (Ha *et al.*, 1995b).

Approches ascendantes Les approches ascendantes partent généralement des pixels et cherchent ensuite à construire des blocs de plus grande dimension. Ces approches sont généralement plus lentes que les approches descendantes. Néanmoins, ces approches sont relativement tolérantes et flexibles. Les méthodes généralement utilisées sont celles basées sur l'utilisation de l'algorithme RLSA (*Run-Length Smoothing Algorithm*) (Wong *et al.*, 1982), l'analyse de spectre (O'Gorman, 1993), la décomposition de Voronoï (Kise *et al.*, 1998) ou encore la détection de lignes de texte (Breuel, 2002).

Approches hybrides Enfin, les approches hybrides regroupent les approches qui combinent la rapidité des approches descendantes avec la robustesse des approches ascendantes. Les méthodes proposées reposent notamment sur l'emploi de filtre de Gabor (Jain et Bhattacharjee, 1992) ou sur l'emploi de signature fractale (Tang *et al.*, 1995).

Shafait *et al.* (2008) proposent une comparaison entre six méthodes de segmentation : une baseline naïve, la recherche de plages blanches, l'algorithme RXYC, l'algorithme RLSA, l'analyse de spectre, la décomposition de Voronoï et, enfin, la détection de lignes de texte. Les auteurs concluent que pour des collections de documents suivant une mise en forme rectangulaire[41] l'analyse de spectre et la décomposition de Voronoï donnent généralement de bons résultats. Pour les collections hétérogènes, la détection de lignes de texte s'avère être la méthode la plus robuste.

2.2.2 Analyse logique

Au sein de la communauté d'Analyse du Document, l'analyse logique consiste à associer la structure géométrique à une structure logique. Cette association est non déterministe (Tang *et al.*, 1996) et s'effectue dans la pratique en (i) étiquetant les unités physiques du document et (ii) en trouvant des relations dites *logiques* entre celles-ci. La structure logique est également définie au sein de l'ISO 8613 (1989) comme :

[39] La notion d'unité atomique est changeante selon les approches et les formats traités.
[40] Tang *et al.* (Tang *et al.*, 1996) parlent d'approches à base de connaissance (pour *Knowledge based*).
[41] Cette mise en page est dite *Manhattan layout*.

> The result of dividing and subdividing the content of a document into increasingly smaller parts, on the basis of the human-perceptible meaning of the content, for example, into chapters, sections, paragraphs.

Selon l'ISO 8613, la structure logique est considérée comme un arbre en constituants dont les nœuds non-terminaux sont des *objets logiques composés*, et les nœuds terminaux sont des *objets logiques basiques*. L'ensemble des étiquettes de ces objets logiques basiques dépend de la tâche traitée (Cattoni *et al.*, 1998). Par exemple pour un corpus de périodique, les étiquettes sont généralement titre, paragraphe, section, etc. Dans le cas d'un corpus épistolaire, d'autres étiquettes logiques peuvent être considérées telles que destinataire, corps de texte, lieu, etc. Les relations entre ces objets sont également dépendantes de la tâche (p. ex. ordre de lecture, lien entre légende et figure, inclusion, etc.) et dépendent d'un modèle de document qui doit être défini *a priori*. Cet état se reflète dans les systèmes commerciaux qui généralement sont restreints à un type de documents (p. ex. chèques, dossier de patients, etc.).

De nombreuses techniques d'analyse logique existent (voir les états de l'art sur le sujet (Mao *et al.*, 2003; Cattoni *et al.*, 1998; Paaß et Konya, 2012)). Nous présentons ici uniquement quatre grandes familles d'approches : la transformation d'arbre, les langages de description, les grammaires formelles et les approches par apprentissage.

Transformation d'arbres Dans cette famille d'approche, les structures géométriques et logiques sont représentées par des arbres. L'analyse logique est alors vu comme un problème de transformation d'arbres. L'un des travaux pionniers est celui de Tsujimoto et Asada (1992). L'arbre géométrique est construit de manière ascendante : les mots sont fusionnés en lignes et, ensuite, celles-ci sont fusionnées en blocs. L'ordonnancement de ces blocs permet la construction de l'arbre géométrique. L'analyse logique est alors effectuée en transformant l'arbre géométriques avec des règles récursives.

Langages de description Les langages de description permettent la description des structures ainsi que l'expression de règles pour les construire. Schürmann *et al.* (1992) proposent le langage FRESCO (*frame representation language for structured documents*). Au sein de celui-ci, les éléments du document sont représentés dans un paradigme orienté objet. Chaque classe d'objets dispose d'attributs et de relations qui la lient aux autres classes. Deux familles de classes sont considérées : les classes physiques (p. ex. document, bloc de texte, ligne, mot, etc.) et les classes logiques (p. ex. titre, adresse, nom d'auteur, date, etc.). L'analyse logique est effectuée au travers d'une stratégie associant les objets de ces deux familles. Cette stratégie repose sur une recherche de solution optimale avec l'algorithme A*.

Grammaires formelles Dans ces approches, l'analyse de document est considérée comme un problème d'analyse syntaxique. Les grammaires utilisées sont essentiellement des grammaires hors-contexte. Krishnamoorthy *et al.* (1993) proposent

une méthode de parsing qui analyse en séquence la structure géométrique et la structure logique[42]. Chaque bloc visuel est traité en (i) le représentant par une chaîne de caractères et (ii) en appliquant une décomposition récursive au travers de grammaires hors-contexte. La décomposition est effectuée en plusieurs étapes. Les symboles non-terminaux d'une étape deviennent les symboles terminaux de l'étape suivante. L'étiquetage est effectué en prenant en compte des contraintes de séquence (p. ex. le nom de l'auteur doit être avant le titre) et de cardinalité (p. ex. deux blocs de colonnes par page au maximum).

Apprentissage artificiel Les approches par apprentissage artificiel cherchent à étendre automatiquement les règles déterministes des approches précédentes ou bien à remplacer ces règles par des décisions probabilistes. Kopec et Chou (1994) proposent d'utiliser des modèles de Markov cachés (HMM) pour extraire les noms et les numéros de téléphone dans des pages jaunes. Esposito *et al.* (1994) proposent d'utiliser de la programmation logique inductive (ILP) à toutes les étapes d'analyse du document[43]. Par la suite, les approches proposées se tournèrent vers des méthodes numériquement plus lourdes. Par exemple, Rangoni et Belaïd (2006) associent aux couches d'un réseau de neurones différents niveaux dans l'analyse du document. La première couche correspond aux traits physiques du document, tandis que les trois couches suivantes correspondent à des représentations de plus en plus fines dans l'analyse logique[44].

Contrairement aux modèles théoriques du TAL présentés en section 2.1, les représentations en Analyse du Document ne sont pas ou peu adaptées à une analyse discursive et permettent difficilement le traitement de structures textuelles qui conjuguent à la fois mise en forme matérielle et phénomènes discursifs (p. ex. structures hiérarchiques, définitions, etc.). Les raisons sont multiples.

Premièrement, un plus grand intérêt est porté à l'analyse géométrique, compliquée pour les documents historiques, les lettres, etc., et non à la construction d'une structure logique en lien avec la structure discursive. Deuxièmement, l'absence de consensus quant aux étiquettes logiques et aux mesures d'évaluation rendent difficile la mise au point de standards pour la structure logique. Paaß et Konya explicitent :

> It is very important to note that in the area of logical layout analysis, there do not exist any standardized benchmarks or evaluation sets, not even algorithms for comparing the results of two different approaches.

[42] « Segmentation and classification must be performed in tandemor, at least, very closely interwoven. » (Krishnamoorthy *et al.*, 1993)

[43] « The success of the machine learning approach to document classification induced us to investigate the possibility of adopting the same approach for the problem of document understanding, that is, for recognizing logical components of a document once it has been classified. » (Esposito *et al.*, 1994)

[44] Cette manière de procéder trouve écho dans les réseaux de neurones dits *profonds*, où chaque couche cachée tente d'apprendre l'abstraction faite par la couche précédente (Bengio, 2009).

Enfin, les frontières entre l'analyse géométrique et l'analyse logique ne sont pas toujours nettes. Cela est notamment vrai pour certains types de documents pour lesquels des informations structurelles sont accessibles préalablement. Par exemple, le résumé d'un article est parfois considéré comme une étiquette logique à part entière. Nous reviendrons sur cette question dans le chapitre 4.

2.3 Formats et structure de document

La distinction classique entre aspect visuel et aspect logique propre au standard ISO 8613 (1989) et à la communauté d'Analyse du Document (section 2.2) se retrouve au sein des formats de documents (*document file format*) proposés ces dernières décennies. Selon le versant visé, deux familles de formats peuvent être distinguées :

- les formats de document reposant sur des **langages de balisage** (*markup languages*) où la structure logique est décrite au travers de balises. Les balises peuvent présenter une sémantique préalablement définie (p. ex. Scribe, LaTeX etc.) ou non (p. ex. SGML, XML, etc.).
- les formats de document reposant sur des **langages de description de page** (*page description languages*) où la structure physique est décrite. Historiquement ces langages visaient à obtenir un rendu visuel identique indépendamment du dispositif d'*impression* (au sens étendu : écran, imprimante, etc.). Ces formats sont soit textuels (p. ex. PCL, Postscript, etc.), soit binaires (p. ex. DVI, PDF, etc.).

Dans la pratique, la situation est néanmoins plus complexe. Par exemple, les années 90 ont vu l'arrivée de formats binaires et fermés mélangeant aspects visuels et logiques tels que le format DOC de Microsoft. C'est pourquoi la confusion est fréquemment faite entre ces types de formats et leurs outils.

Dans cette section, nous limitons notre propos à deux situations particulières en lien avec ces familles de formats. Dans un premier temps, nous regardons dans quelle mesure les langages de balisage sont adéquats pour représenter la structure logique des documents. Dans un second temps, nous présentons quelques réflexions sur l'intégration de la structure logique au sein des langages de description de page.

2.3.1 Langages de balisage

Dans cette partie, nous discutons le compromis entre la structure visuelle et la structure logique au travers de la description brève de trois cas concernant respectivement le système de composition TeX, les formats LaTeX et HTML, et enfin le format XML.

Le système de composition TeX À proprement parler, TeX n'est ni un langage, ni un format. Dans l'ouvrage *TeXbook*, Knuth le définit comme un système de composition dédié à la préparation de documents techniques et mathématiques (1984) :

> This is a handbook about TEX a new typesetting system intended for the creation of beautiful books—and especially for books that contain a lot of mathematics. By preparing a manuscript in TEX format, you will be telling a computer exactly how the manuscript is to be transformed into pages whose typographic quality is comparable to that of the world's finest printers ; yet you won't need to do much more work than would be involved if you were simply typing the manuscript on an ordinary typewriter.

TEX offre un contrôle fin des structures logiques et visuelles au travers d'environ 900 séquences de contrôle. Environ 300 d'entre elles sont appelées primitives et ne peuvent être décomposées en fonctions TEX de plus bas niveau. Par exemple, `\input{}` permet l'inclusion d'un fichier externe de commandes. Les 600 autres séquences de contrôle sont proposées au travers du format Plain TEX, fourni nativement avec TEX[45]. Celles-ci définissent essentiellement des éléments de niveau visuel (Schrod, 1991). Par exemple, `\rightline{}` permet l'alignement à droite du contenu donné en argument. Par la suite, Lamport proposera le format LATEX dont les séquences de contrôle apporteront un traitement plus fin de la structure logique.

Notons que le rendu graphique du Plain TEX est traditionnellement obtenu par sa compilation en format DVI (*Device Independent Format*). Dans ce format binaire, chacune des commandes désigne une opération visuelle de bas niveau (Knuth, 1995), rendant le rendu visuel indépendant des plates-formes.

Les formats LATEX et HTML Sur la base du format Plain TEX, le format LATEX fut proposé par Lamport en 1984[46] (Lamport, 1994). Ce format est très rapidement devenu une « lingua franca » dans de nombreuses disciplines scientifiques. Ce succès s'explique notamment par l'intérêt donné aux structures logiques, laissant de côté la partie associée à la mise en forme. Lamport (1994) explique clairement :

> The primary function of almost all the LATEX commands that you type should be to describe the logical structure of your document. As you are writing your document, you should be concerned whith its logical structure, not its visual appearance. The LATEX approach to typesetting can therefore be characterized as *logical design*. (...) LATEX was designed to free you from formatting concerns, allowing you to concentrate on writing.

LATEX étant fondé sur TEX, il est également possible d'obtenir un rendu visuel au travers du format DVI. À l'heure actuelle, le format PDF est généralement préféré et peut être obtenu avec l'outil `pdftex`[47].

[45] Knuth souligne que ce format est un exemple parmi d'autres : « However, you should keep in mind that plain TEX is only one of countless formats that can be designed on top of TEX's primitives ; if you want some other format, it will usually be possible to adapt TEX so that it will handle whatever you have in mind. » (Knuth, 1984, p. 11)

[46] Historiquement LATEX fut implémenté par Lamport jusqu'à la version 2.09. À partir de 1994, un groupe s'occupa des développements ultérieurs, appelés LATEX 2_ε. L'usage a gardé le terme LATEX.

[47] http://www.tug.org/applications/pdftex/

Le HTML (*HyperText Markup Language*) est un format textuel permettant la représentation, au travers de balises, de pages Web. Syntaxiquement, le HTML hérite du format SGML (*Standard Generalized Markup Language*) auquel il ajoute néanmoins une sémantique prédéfinie sur un ensemble fini de balises (Berners-Lee et Connolly, 1993). Une distinction est faite entre les *styles logiques* et les *styles physiques* qu'il est possible d'associer aux caractères. Les premiers concernent l'intention communicative (ce *mot* est mis en emphase), tandis que les seconds concernent la représentation physique (ce *mot* est en italique). La difficulté est que les styles logiques trouvent nécessairement une traduction physique. Berners-Lee et Connolly explicitent :

> Some (...) styles are more explicit than others about how they should be physically represented. The logical styles should be used wherever possible, unless for example it is necessary to refer to the formatting in the text. (Eg, "The italic parts are mandatory".)

Cette préférence pour les styles logiques s'explique notamment par l'absence d'un standard de représentation visuelle pour le format HTML. Chaque moteur de rendu HTML (*web browser engine*) dispose de ses propres conventions. Il n'est alors pas assuré que l'application d'un style physique soit faite telle que visuellement attendue par son auteur[48]. *A contrario*, les styles logiques gardent une cohérence du document de manière relativement indépendante des procédés liés à leur génération visuelle. Dans ce contexte, l'intérêt est porté sur l'emphase, le contraste et l'alternance.

Malgré l'objectif commun des formats LaTeX et HTML visant à séparer le contenu de sa mise en forme[49], il n'est néanmoins pas possible de les considérer comme l'expression stricte d'une structure logique abstraite. Deux raisons sont évoquées ici :

- Le mélange des balises visuelles et logiques (par héritage pour le LaTeX ou par choix pour le HTML) conduit à un enchevêtrement d'indices[50]. Par exemple, dans la table 2.3, nous montrons pour certains styles logiques que des pendants visuels peuvent leur être visuellement substitués. Cela est obligatoire pour l'emphase forte en LaTeX pour laquelle il n'existe pas nativement de contrepartie logique.

  Néanmoins notons qu'un objectif de « *sémantisation* » est poursuivi par le W3C ces dernières années (Berners-Lee *et al.*, 2001). Par exemple, la balise HTML `<tt>` n'est plus supportée dans la spécification actuelle du HTML (Hickson *et al.*, 2014)

[48] « Browsers unable to display a specified style may render it in some alternative, or the default, style, with some loss of quality for the reader. Some implementations may ignore these tags altogether, so information providers should attempt not to rely on them as essential to the information content. » (Berners-Lee et Connolly, 1993)

[49] « The common philosophy of these languages is that markup should abstract from the visual appearance of the document, using concepts like paragraph that might be realized graphically in different ways, depending on a separate style definition. » (Power *et al.*, 2003)

[50] Schrod (1991) le souligne à propos du LaTeX : « This results in the effect that the author can use a mixture of structural information and explicit layout information – a situation with a high potency of features that nevertheless can (and does) lead to a lot of typographic nonsense ! »

qui, pour l'expression d'un programme informatique, lui préfère `<code>`, `<samp>` (partie de code), `<kbd>` (entrée clavier) ou `<var>` (variable). À un niveau plus élevé, cette spécification préconise également le remplacement du container `<div>` par un pendant logique (dit *sémantique*) tel que `<section>`, `<article>`, etc.

Logique	HTML	LaTeX	Visuel	HTML	LaTeX
emphase	`<em>`	`\emph{ }`	italique	`<i>`	`\textit{ }`
emphase forte	`<strong>`	-	gras	`<b>`	`\textbf{ }`
code	`<code>`	`\verb| |`	chasse fixe	`<tt>`	`\texttt{ }`

TABLE 2.3 : Exemples de correspondances entre balises logiques et visuelles dans les formats LaTeX et HTML

- Certains éléments appartenant à la structure logique sont signalés par la mise en forme au sein des formats LaTeX et HTML, et non au travers de balises dédiées. Au niveau du bloc textuel, l'auteur désirant signaler un paragraphe, ou plus généralement un alinéa, en LaTeX doit le faire au moyen de dispositifs dispositionnels (p. ex. une ligne vide avant, et un retour à la ligne après). Ceux-ci ont alors un rôle proche du délimiteur de Nunberg (section 2.1.1).

  À un niveau plus fin, Power *et al.* (2003) font remarquer que les formats LaTeX et HTML ne permettent pas de signaler les phrases et les clauses autrement que par la ponctuation. Ceci entre en contradiction avec le fait que les phrases et les clauses puissent être exprimées au travers d'une large variété de ponctuation sans que leur statut abstrait n'en soit changé. Par exemple, une même clause peut être terminée par un point-virgule (comme le recommande Nunberg), par des points de suspension (comme le fait Céline), par des tirets (comme le fait Flaubert) ou enfin par l'absence de ponctuation (comme le fait Joyce).

Le format XML Le format XML (*eXtensible Markup Language*) est un sous-ensemble du SGML pour lequel les contraintes syntaxiques sont plus strictes[51]. Dans la spécification XML 1.0 (Bray *et al.*, 1998), une distinction est faite entre la structure physique et la structure logique. Toutes les deux peuvent être définies au sein d'une DTD (*Document Type Definition*). Une DTD permet la définition d'une classe de documents au travers d'une grammaire hors-contexte. Dans les spécifications du W3C, celle-ci est décrite au travers de la notation Backus-Naur (BNF). Chaque règle définit un symbole comme suit :

```
symbole ::= expression
```

[51] Ces contraintes concernent notamment l'obligation de fermer les balises ouvertes, spécifier les valeurs des attributs commes des littéraux, etc. Se référer à la spécification relative à la comparaison entre XML et SGML (Clark, 1997).

Un document XML est correctement formé (*well-formed*) s'il est syntaxiquement correct, et est valide si ses structures physique et logique valident la DTD. Dans la suite, nous décrivons les unités et l'expression des contraintes syntaxiques pour ces structures :

- La structure physique concerne ici les différentes unités physiques composant le document XML en tant que tel. L'intérêt pour cette structure trouve son origine dans le fait que les documents XML ne sont pas obligatoirement écrits sur le disque, mais peuvent être construits dynamiquement et échangés au travers de requêtes entre applications. La structure physique est composée d'unités de stockage (*storage units*) appelées *entités*. Les entités peuvent contenir des flux de natures différentes (p. ex. données textuelles, binaires, ou encore XML). Avant d'être utilisées, les entités doivent être déclarées[52]. La règle de déclaration d'une entité est :

  ```
  EntityDecl ::= '<!ENTITY' S Name S EntityDef S? '>'
  ```

  où `S` est un ou plusieurs espaces, `Name` est le nom donné à l'entité et `EntityDef` désigne le contenu et la nature l'entité. Les caractères entre apostrophes sont des littéraux. L'exemple ci-dessous est la déclaration d'une image avec un chemin relatif comme entité externe :

  ```
  <!ENTITY chat SYSTEM "./images/chat.gif" NDATA gif >
  ```

- La structure logique concerne les règles de composition des éléments abstraits du document. Les unités utilisées dans cette structure sont appelées *éléments*. Ceux-ci ont un type unique et un ensemble d'attributs. Formellement, la déclaration d'un élément est défini par la règle suivante :

  ```
  ElementDecl ::= '<!ELEMENT' S Name S contentspec S? '>'
  ```

  où `Name` est le nom donné à l'élément et `contentspec` exprime les contraintes sur le contenu de l'élément déclaré. Par exemple, la déclaration suivante contraint les éléments de type SECTION à être composés d'éléments TITRE ou PARAGRAPHE :

  ```
  <!ELEMENT SECTION (TITRE|PARAGRAPHE)>
  ```

Par l'expressivité offerte par la DTD, le format XML est le plus à même de permettre une distinction nette entre structure physique et structure logique. Plusieurs formats dérivés du XML ont proposé leur propre DTD dans cet objectif. Par exemple, les formats DocBook[53] et LinuxDoc[54] proposent au travers de balises telles que la section, le paragraphe, les listes à puces, ou encore l'emphase de décrire uniquement le versant logique des documents[55]. Ce choix leur permet d'être très facilement transcrits dans d'autres formats. Néanmoins, notons que ces formats restent peu utilisés. Deux raisons peuvent expliquer cela : (i) il est peu ergonomique d'éditer des balises XML et (ii) ces formats ne supportent pas directement le rendu de formules.

[52] Notons que l'entité document ne nécessite pas d'être définie préalablement pour être utilisée. Cette dernière est la racine de l'arbre et contient récurisvement les entités du document XML.
[53] http://docbook.org/
[54] http://www.tldp.org/
[55] Notons que ces formats étendaient originellement SGML.

2.3.2 Langages de description de page

Dans cette partie, nous décrivons le compromis entre les aspects visuels et logiques au travers de deux cas : le langage PostScript et le format PDF.

Le langage PostScript Le PostScript est un langage de description de page proposé par Adobe en 1982. Celui-ci permet l'encodage et le rendu des éléments constitutifs de la structure visuelle d'un document (p. ex. texte, fontes, image, couleurs, etc.) (Adobe, 1985). L'encodage est effectué au travers d'instructions textuelles décrivant des primitives (p. ex. lignes, formes, caractères[56], etc.). Cette représentation vectorielle est donnée à un interpréteur RIP (*Raster Image Processor*) pour obtenir un rendu matriciel. Nous donnons en figure 2.7 un exemple d'instructions PostScript et de leur rendu visuel.

```
%!PS
/Courier
100 selectfont
10 10 moveto
(abc) show
showpage
```

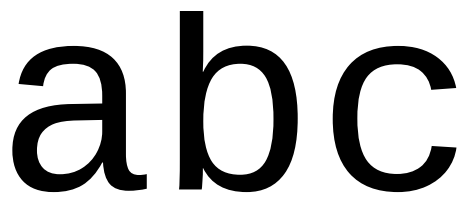

FIGURE 2.7 : Exemple d'instructions PostScript et de leur rendu visuel

Ce mécanisme d'interprétation permet d'obtenir une structure visuelle identique sur une imprimante ou un écran d'ordinateur disposant d'un interpréteur adéquat. Cela permet également de décrire des documents sans que cela nécessite de stocker de larges matrices de pixels. Cet objectif va dans la direction opposée de l'expression d'une structure logique. Dans le langage PostScript, celle-ci n'est pas exprimée (Adobe, 1992b).

Pour faciliter l'échange de documents écrits en PostScript, Adobe a proposé le format EPS (*Encapsulated PostScript*) (Adobe, 1992a). Ce format se présente sous la forme d'une série d'instructions PostScript (avec certaines restrictions syntaxiques) qui décrivent une page unique. Le EPS peut être inclus (« encapsulé ») au sein d'un autre document. Pour des raisons historiques, le EPS offre également la possibilité de contenir une représentation matricielle de son contenu, utile pour les ordinateurs ne disposant pas d'interpréteur PostScript ou n'ayant pas de ressources suffisantes.

Le format PDF Le PDF (*Portable Document Format*) a été proposé par la société Adobe en 1993. La naissance de ce format s'explique par les limites techniques des ordinateurs de l'époque pour interpréter de larges et complexes documents PostScript. Warnock (1991), auteur du projet PDF, explique :

> The Display PostScript and PostScript solutions are the correct long-term solution as the power of machines increases over time, but this solution offers little help for the vast majority of today's users with today's machines.[57]

[56] Les caractères sont exprimés au travers de la fonte PostScript (dite *PostScript Type 1*). Celle-ci s'opposa à la TrueType de Apple dans les années 90 (Phinney, 2004).

[57] Le *Display PostScript* est un moteur qui étend les capacités d'affichage du PostScript sur un écran.

Le format PDF hérite de l'objectif du format EPS visant le partage des documents[58], mais propose des améliorations conséquentes telles qu'un interpréteur léger, la possibilité de décrire plusieurs pages, des fonctionnalités de recherche textuelle, etc. Contrairement au mécanisme d'accès séquentiel du PostScript, le format PDF permet un accès direct aux objets qu'il contient grâce à une structure en arbre.

En effet, la structure d'un document PDF est représentée sous la forme d'une hiérarchie ordonnant des objets de différents types (chaînes de caractères, nombres entiers, tableaux, flux, dictionnaires, etc.) (Adobe, 2008, section 7.7.1). À la racine de l'arbre, un dictionnaire, appelé catalogue, contient les références vers les nœuds immédiatement inférieurs. Ce catalogue peut, par exemple, être déclaré tel qu'en figure 2.8.

```
1 0 obj
<< /Type /Catalog
 /Outlines 2 0 R
 /Pages 3 0 R
>>
```

FIGURE 2.8 : Exemple de déclaration du catalogue dans un document PDF

Le premier chiffre est l'identifiant de l'objet et le second est le numéro de révision. La chaîne de caractères `"obj"` et les crochets indiquent qu'il s'agit d'un objet de la classe dictionnaire. Trois paires clef-valeur sont définies. La première définit le dictionnaire comme un catalogue. La seconde est une référence vers l'objet d'identifiant 2 qui est la racine du sous-arbre représentant le sommaire du document (*outline hierarchy*). Enfin, la dernière est une référence vers l'objet d'identifiant 3 qui constitue la racine d'un sous-arbre des pages du document (*page tree*). En figure 2.9, nous donnons un exemple de la hiérarchie d'un PDF. Notons que d'autres types d'objets peuvent être ajoutés.

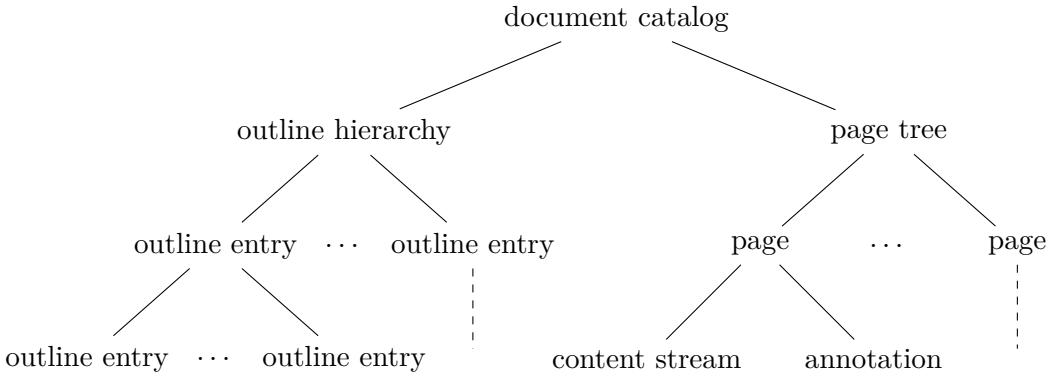

FIGURE 2.9 : Schéma simplifié de la hiérarchie d'un PDF

[58] « Imagine being able to send full text and graphics documents (newspapers, magazine articles, technical manuals etc.) over electronic mail distribution networks. These documents could be viewed on any machine and any selected document could be printed locally. » (Warnock, 1991)

Le sous-arbre du sommaire (*outline hierarchy*) permet à un utilisateur de se déplacer dans l'ensemble du document. À chacun des nœuds, appelés signets (*outline entry*), est associé une destination physique dans le document (Adobe, 2008, section 12.3.3). Ce sous-arbre est généralement créé à la volée à partir de la titraille du document original.

Le sous-arbre des pages (*page tree*) définit l'ordre des pages dans le document. Les nœuds terminaux sont des objets de type page et les nœuds non-terminaux sont les objets sur ces pages telles que les annotations sur le document (*annotation*), les données (*content stream*), etc. Notons que l'ordre de cet arbre n'est pas nécessairement relatif à l'ordre de lecture (et à la structure logique du document) (Adobe, 2008, section 7.7.3.2).

La spécification PDF 1.3 (Adobe, 2000) est la première qui ajoute une description de la structure logique. La déclaration de cette structure logique est proche de celles des langages à balises tels que le HTML et XML. Celle-ci prend la forme d'un arbre (*structure tree*) attaché au catalogue. Deux types de nœuds sont distingués : les nœuds non-terminaux qui sont des éléments structurels (*structure elements*) et les nœuds terminaux qui sont des items (*content items*). Les éléments structurels ont chacun une étiquette (p. ex. chapitre, paragraphe, liste, etc.). Adobe a proposé une liste d'étiquettes. Les items associés aux éléments structurels sont des références vers le contenu (p. ex. caractères, images, etc.). La figure 2.10 propose une hiérarchie de PDF avec une structure logique.

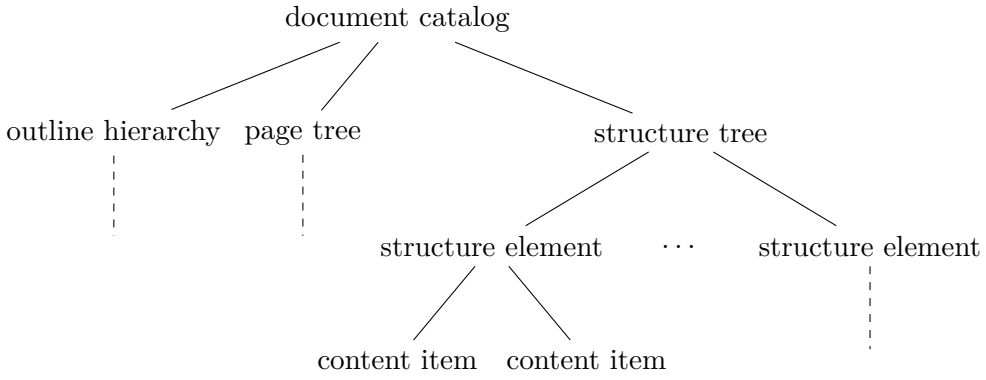

FIGURE 2.10 : Schéma simplifié de la hiérarchie d'un PDF avec sa structure logique

L'ajout d'une structure logique dans un format initialement prévu pour la description graphique des pages témoigne de l'intérêt porté pour celle-ci. En effet, l'adoption massive du PDF comme document d'échange et d'archivage à fait réaliser la nécessité pratique d'une structure abstraite du document. Celle-ci ouvre la voie à des applications avancées telle qu'une lecture adaptée pour des personnes malvoyantes ou encore la remise en forme (*reflowing*) pour des écrans de petite taille (Hardy et Brailsford, 2002).

Néanmoins, deux difficultés empêchent une exploitation de cette structure logique. Premièrement, les documents PDF ne sont majoritairement pas étiquetés logiquement. Seuls certains programmes propriétaires permettent d'associer des étiquettes et un coût manuel lourd est généralement associé à cette procédure d'étiquetage logique.

Deuxièmement, il n'y a pas de consensus quant aux étiquettes. La spécification PDF 1.4 (2001) proposera sous l'appellation de *Tagged PDF* un ensemble standard d'étiquettes (Adobe, 2001, section 9.7.4). Néanmoins, cet ensemble est complexe (plus de 60 étiquettes) et inconsistant (p. ex. présence de la formule, mais pas de théorème ou de définition).

2.4 Discussion

Dans ce chapitre nous avons vu que la structure de document est un sujet complexe et qu'elle est traitée différemment selon les communautés scientifiques. Notre propos s'est centré sur la situation au sein de deux communautés spécifiques.

Au sein de la communauté de Traitement Automatique du Langage, l'accent est mis sur le passage entre la structure discursive et la structure logique. Dans ce contexte, les modèles proposés sont difficilement adaptables à un processus d'analyse à partir de la structure visuelle. Inversement, au sein de la communauté d'Analyse du Document, l'intérêt est porté sur le lien entre la structure visuelle et la structure logique des documents. Cependant, aucune réflexion n'est faite sur la représentation hiérarchique de la structure logique, ses frontières et son adéquation avec la structure rhétorique.

Cette situation se reflète dans l'évolution du format PDF qui a été conçu originellement pour représenter la structure visuelle, mais qui est actuellement étendu pour représenter la structure logique en vue d'améliorer des traitements textuels de haut niveau. Or, paradoxalement, il apparaît que la structure logique est encore mal définie et ses frontières restent peu étudiées.

D'un côté, bien que la distinction entre la structure physique et la structure logique paraisse immédiate, elle n'est néanmoins pas assurée dans la pratique. La profusion massive d'outils WYSISYG (*What You See Is What You Get*) a impliqué chez de nombreux utilisateurs l'adoption d'une conception avant tout visuelle. Lamport (1994) explicite :

> Wysiwyg programs replace LaTeX's logical design with *visual design*. Visual design is fine for short, simple documents like letters and memos. It is not good for more complex documents such as scientific papers.

À cela s'ajoute le fait que ces outils proposent (ou obligent) l'utilisation d'un format associé. Pour des utilisateurs non-experts, il peut alors devenir difficile de faire la distinction entre à la fois le format, la structure physique et la structure logique du document[59]. Plus malencontreux et indépendamment de l'outil, les contraintes de mise en forme interfèrent avec la structure logique. Par exemple, dans le cas des documents numériques où la notion de page est utilisée, les pratiques typographiques veulent que les lignes veuves et orphelines soient évitées. Également, pour les documents à balises sur le web, il n'est pas rare de voir que les auteurs favorisent l'aspect visuel au détriment de la syntaxe.

[59] Parfois, la distinction n'est également pas faite entre le format et l'outil lui-même.

De l'autre côté, à un niveau plus abstrait, il apparaît difficile de distinguer nettement ce qui tient du logique de ce qui tient du discursif. Si nous admettons ici que la structure rhétorique est indépendante du médium, alors il est important de considérer les étiquettes telles que section, paragraphe, item, etc. comme des catégories linguistiques plutôt que rhétoriques. Un même message pourrait être donné oralement (p. ex. présentation) ou bien par écrit (p. ex. rapport). La structure logique est donc bien une *structure de document*. Cependant, il semble difficile de séparer catégoriquement les deux structures, car l'interprétation humaine est immédiate. Cela est notamment dû au fait que nous nous référons aux segments du document en utilisant le nom qui décrit leur rôle rhétorique (p. ex. résumé, introduction, conclusion, etc.).

Notre travail se situe dans la suite des modèles théoriques de structure de document proposés en TAL. Toutefois, nous soulignons la nécessité d'avoir un modèle qui puisse être adapté à un processus automatique d'analyse. Cela est nécessaire pour envisager une perspective d'extraction de relations en s'aidant des éléments de mise en forme.

C'est pourquoi, en partie II, nous proposons notre modèle et nous l'implémentons. Celui-ci sera intégré, en partie III, à l'extraction de relations à partir de structures énumératives. Ces structures textuelles sont intéressantes, car elles présentent un terrain propice aux relations hiérarchiques, et leur mise en forme laisse envisager leur repérage automatique. Le chapitre suivant sera consacré à un état de l'art sur ces structures textuelles.

Chapitre 3

Structures énumératives

Sommaire

Dans ce chapitre, nous nous intéressons aux structures énumératives (SE). Du point de vue sémantique, ces structures textuelles sont intéressantes, car elles sont propices aux relations sémantiques hiérarchiques, utiles à la création de ressources. Du point de vue de leur réalisation, les SE mettent en œuvre différents mécanismes : elles peuvent passer de formes linéaires, réalisées au travers de constructions syntaxiques (juxtaposition, coordination, etc.), à des formes matérialisées typographiquement et dispositionnellement qui les rendent perceptibles à la surface des textes. Ce marquage visuel permet d'envisager plus aisément leur identification dans les textes, mais également le bornage de leurs composants internes.

La division de ce chapitre est la suivante. Dans un premier temps, nous présentons les problèmes liés à la définition et à la délimitation des SE. Dans un second temps, nous présentons deux typologies des SE sur lesquelles nous nous appuierons. Enfin, nous montrons des travaux d'extraction de connaissances qui exploitent les SE.

3.1 Définition et délimitation des structures énumératives

Les travaux dans la littérature divergent dans leur appréhension respective des SE. Cette section discute les questions de la définition et de la délimitation des SE.

3.1.1 Problème de la définition

Les approches sur les SE sont nombreuses et variées, et il n'existe pas de consensus quant à la manière de définir ces structures. Dans cette section, nous nous intéressons aux définitions données SE en présentant quelques études :

- Une des premières études fut celle de Turco et Coltier (1988). Les auteurs considèrent comme *énumérations* des structures introduites par des marqueurs les rendant facilement identifiables par le lecteur (ou par l'interlocuteur dans le cas de l'oral). Ces marqueurs sont assimilés à des marqueurs d'intégration linéaires (MIL) (p. ex. *D'une part, D'autre part, Premièrement, Deuxièmement,* etc.) et « accompagnent l'énumération sans fournir de précision autre que le fait que le segment discursif qu'ils introduisent est à intégrer de façon linéaire dans la série » (Turco et Coltier, 1988, p. 57)[1]. Ainsi, l'énumération est considérée avant tout comme une progression linéaire perceptible.

- Adam et Revaz (1989) ont prolongé les travaux de Turco et Coltier (1988) et s'intéressent à la notion d'ordre en mettant l'accent sur les dimensions de temporelle et spatiale. Plusieurs types de marqueurs sont étudiés par les auteurs : les organisateurs énumératifs (qui correspondent aux marqueurs de Turco et Coltier[2]), les organisateurs temporels (p. ex. *la veille, le lendemain*, etc.) et les organisateurs spatiaux (p. ex. *au nord, au sud, à gauche, à droite*, etc.). Dans ce contexte, les auteurs définissent l'*énumération* comme une structure où l'ordre n'intervient pas (Adam et Revaz, 1989, p. 66) :

  > L'énumération (de parties, de propriétés ou d'actions) est une des opérations descriptives les plus élémentaires. Dans tous les cas, il s'agit de développer linéairement un ensemble de propositions dont l'organisation n'est à l'origine ni causale (argumentation), ni chronologique (narration ou injonction-instruction de la recette ou de la notice de montage). A priori, une énumération n'est régie par aucun ordre.

- Hovy et Arens (1991) mettent également en avant la notion d'ordre spatial et temporel. Les auteurs parlent de dispositifs textuels (*textual devices*) et distinguent les listes d'items (*itemization*) où l'ordre n'intervient pas des listes énumérées (*enumerations*) qui prennent en compte l'ordre des composants :

[1] Également cité par Adam et Revaz (1989).

[2] « A la suite de G. Turco & D. Coltier (1988), nous considérons les organisateurs énumératifs comme des marqueurs d'intégration linéaire (dorénavant M.I.L.) » (Adam et Revaz, 1989, p. 66)

> (...) the text structure relation SEQUENCE can generally be formatted as an enumerated list. The enumeration follows the sequence of the relation, which is planned in expression of some underlying semantic ordering of the items involved, for example, time, location, etc.

- Damamme-Gilbert (1989) utilise le terme de *série énumérative*. La série énumérative est avant tout définie selon des critères syntaxiques stricts, avant d'être un motif stylistique ou sémantique :

  > (...) expression linguistique formée d'un nombre minimum de trois termes (mots, syntagmes, unités d'énoncé) qui appartiennent à des catégories morphologiques ou grammaticales identiques ou équivalentes, qui occupent une fonction identique dans la syntaxe de l'énoncé et qui, placées côte à côte, sont coordonnées ou reliées par un signe de ponctuation.

  Dans ce cadre, c'est le parallélisme syntaxique qui est mis en avant.

- Jackiewicz et Minel (2003) s'intéressent à ce qu'ils appellent des *séries*. Celles-ci sont des structures textuelles linéaires organisées par le discours. Ce travail s'inscrit dans la notion de *cadre de discours* de Charolles (1997). Les cadres amènent des liens de cohésion qui guident l'interprétation du contenu propositionnel[3]. Dans ce contexte, Jackiewicz et Minel considèrent les MIL comme des marques cadratives et répertorient ceux qui occurrent fréquemment dans les séries. Les séries sont alors un moyen d'obtenir une unité structurante plus grande que la phrase, mais plus étroite que la section[4]. Hernandez et Grau (2005) parlent de « structures fines ». Ce travail de structuration a ouvert la voie au développement de systèmes d'aide à la navigation prenant en compte les aspects discursifs (Couto *et al.*, 2004).

- Les travaux de Pascual et Virbel ont proposé un socle théorique en lien avec le MAT (section 2.1.3). Pour Pascual (1991), « énumérer, c'est conférer une égalité d'importance à un ensemble d'objets, et ensuite c'est ordonner ces objets selon des critères variés. Lorsqu'on réalise une énumération, le plus important est de manifester notre intention d'énumérer. » Dans ce cas, le motif mis en avant est sémantique, — intentionnel dans le MAT —, avant d'être syntaxique. Virbel généralise davantage la définition afin de prendre en compte des phénomènes où les objets énumérés ne sont pas visuellement ou fonctionnellement équivalents[5]. Virbel (1999) définit l'énumération comme :

[3] L'interprétation est comprise comme l'identification de la cohérence. La distinction classique *cohérence - cohésion* est faite ici (Halliday et Hasan, 1976).

[4] « La prise en considération des indicateurs graphiques et typographiques constitue un préalable, mais les phrases se révèlent bien souvent des unités trop étroites et les paragraphes ou les sections, des unités trop vastes. » (Jackiewicz et Minel, 2003)

[5] « Les cas les plus courant d'énumération, et les seuls à notre connaissance qui soient évoqués dans la littérature, sont ceux où les items constituent des segments qui possèdent une identité de fonction (fonction syntaxique au sein d'une phrase, ou textuelle au sein d'un texte) : l'énumération vise alors à rendre plus manifeste une structure que l'on peut globalement caractériser comme étant de type coordinatif. » (Virbel, 1999)

> l'acte textuel qui consiste à transposer textuellement la coénumérabilité des entités recensées par la coénumérabilité des segments linguistiques qui les décrivent, ceux-ci devenant par le fait les entités constitutives de l'énumération (les items).

Ainsi, définir si une structure textuelle est une SE revient à vérifier si ses constituants sont coénumérés. Cette coénumération dépend étroitement de l'interprétation des marques de mise en forme, mais aussi de la structure rhétorique.

Dans ce travail, nous nous sommes appuyés sur la définition donnée par Virbel, ainsi que la terminologie associée. Dans ce cadre, une SE est un objet textuel (Section 2.1.3) composé d'une amorce et d'une énumération, elle-même composée d'items. Nous donnons les définitions de ces objets ci-dessous (Luc, 2000, p.102) :

- une **amorce** est une définie comme une phrase introduisant l'énumération.
- une **énumération** est définie comme un ensemble d'au moins deux items.
- un **item** est défini comme une entité coénumérée. Ces items peuvent être de granularité variable (de la clause textuelle au paragraphe).

Facultativement, une conclusion, appelée **clôture**, peut être ajoutée à la fin.

De nombreux travaux adoptèrent également cette terminologie (Péry-Woodley, 2000; Ho-Dac *et al.*, 2004; Porhiel, 2007; Bras *et al.*, 2008; Vergez-Couret *et al.*, 2011; Laignelet *et al.*, 2011). Néanmoins, par son critère unique de coénumérabilité, la définition de Virbel pose intrinsèquement le problème de la délimitation.

3.1.2 Problème de la délimitation

La question de la délimitation suit directement celle de la définition, et est inhérente à l'étude en corpus. À partir de quel moment une structure hiérarchique est-elle une SE ? Est-il possible de parler de SE en présence de coordination ? Certaines positions prônent des critères syntaxiques ou dispositionnels stricts (Damamme-Gilbert, 1989). D'autres positions mettent davantage en avant des critères interprétatifs. Par exemple, Ho-Dac (2007) suggère l'hypothèse que l'organisation discursive puisse être représentée par une structure énumérative globale.

Virbel (1999) propose un recueil avec une série d'exemples questionnant l'unité des . L'auteur explicite sa mise en pratique :

> (...) la définition de l'énumération pose entre autres des problèmes de délimitation, et ceux-ci ne peuvent être abordés que conjointement avec d'autres (les séries, les structures en divisions, la numérotation d'objets textuels, au moins).

Afin de déterminer le comportement de la SE, Virbel propose un panel de SE présentées avec leur co-texte immédiat et organisées selon leur déviation respective par rapport à une SE idéale dont les items sont fonctionnellement et visuellement équivalents. Ces

déviations permettent alors de montrer l'existence d'un continuum dans l'expression des SE entre, d'un côté des indices de mise en forme purement lexicaux et syntaxiques, et d'un autre côté des indices typographiques et dispositionnels[6] :

- Les SE réalisées majoritairement par des indices lexicaux et syntaxiques sont appelées SE linéaires (ou horizontales). À leur niveau le plus fin, celles-ci peuvent présenter des items intra-phrastiques (exemple (3.a) issu de (Virbel, 1999)).
- Les SE majoritairement réalisées par des indices typographiques et dispositionnels sont appelées SE verticales. Celles-ci posent la question de la borne supérieure où des items paragraphiques (exemple (3.b) issu de (Virbel, 1999)) ou ultra-paragraphiques (p. ex. les parties d'un chapitre) peuvent être rencontrés.

(3.a) Luc vend et achète des meubles

(3.b)

Cette méthode présentait toutefois quelques inconvénients :

- COMPOSE ne permettant pratiquement pas de traiter de formules mathématiques, certaines équipes préféraient utiliser d'autres systèmes de traitement de texte et fournissaient des prêts-à-clicher qui, bien que respectant la maquette globale, donnaient une impression de non-homogénéité.
- COMPOSE n'était utilisable au mieux qu'avec une imprimante à marguerite, donc avec une seule police de caractères à chasse fixe. Il n'y avait donc pas la moindre possibilité typographique autre que le souligné.
- Certains services, ne disposant pas d'accès à Multics, fournissaient un texte tapé à la machine à écrire, ce qui augmentait l'hétérogénéité du document.
- Un certain nombre de choses globales au document (index, gestion des espaces pour figures etc.) n'étaient que difficilement utilisables.
- etc.

Orthogonalement, d'autres variations peuvent apparaître. Virbel présente des exemples où apparaissent notamment des asymétries syntaxiques, des variations de coordinateurs, des imbrications, des entrelacements, etc. Ces exemples questionnent « l'unité du type » même et soulignent la nécessité d'une interprétation presque au cas par cas.

Dans ce cadre, l'établissement d'une typologie complète n'est pas envisageable. Seule la circonscription du problème à un domaine et à une série d'objectifs préalablement définis permet d'envisager l'établissement de classifications.

[6] Ce balancement entre lexico-syntaxique et typo-dispositionnel a déjà été discuté lors de la présentation de la Mise en Forme Matérielle et de l'évocation de l'objet textuel de la Définition (section 2.1.3).

3.2 Typologies des structures énumératives

Dans cette section, nous présentons deux typologies de SE sur lesquelles notre travail va s'appuyer. La première typologie s'inscrit dans le travail de Luc (2000). La seconde typologie est proposée par Ho-Dac, Péry-Woodley et Tanguy (2010).

D'autres typologies ont été proposées, telles que celle de Porhiel (2007) et Vergez-Couret *et al.* (2008) où les SE à un temps sont opposées aux SE à deux temps, mais ne sont pas évoquées ici.

3.2.1 Typologie de Luc (2000)

Dans son travail, Luc (2000) propose une typologie des SE. Cette typologie s'inscrit dans son travail de composition du MAT et de la RST. La SE est alors considérée comme un lieu privilégié pour l'étude des interactions entre les marques visuelles et les marques de cohésion discursive. Dans cette section, nous exposons :

- la typologie mise en place par Luc ;
- deux exemples de SE analysées selon cette typologie.

Typologie des énumérations Sur la base d'une étude en corpus, Luc identifie trois catégories de problèmes liés aux énumérations au sein des SE. Ceux-ci concernent respectivement (i) les énumérations dont les items entretiennent des relations entre eux, (ii) les énumérations dont les items ne sont pas visuellement équivalents et (iii) les énumérations dont les items entretiennent des relations avec des objets n'appartenant pas à la SE. Au sein de ces catégories orthogonales, Luc propose plusieurs types d'énumérations. Pour chaque catégorie donnée, une énumération ne peut appartenir qu'à un seul type.

Première catégorie

énumération syntagmatique est une énumération dont les items présentent des relations de dépendance (syntaxique ou rhétorique) successives.

énumération paradigmatique est une énumération dont les items sont fonctionnellement équivalents (syntaxiquement ou rhétoriquement).

énumération hybride est une énumération dont au moins deux items sont fonctionnellement équivalents et dont un item est dépendant d'un autre item.

Seconde catégorie

énumération visuellement homogène est une énumération dont tous les items sont visuellement équivalents.

énumération visuellement hétérogène est une énumération dont au moins un item est visuellement différent des autres items.

Troisième catégorie

énumération liée est une énumération au sein de laquelle au moins un item rencontre l'une des conditions suivantes :

- il entretient une relation avec un objet textuel externe à la SE.
- il contient une autre énumération imbriquée.

énumération isolée est une énumération dont les items n'entretiennent aucune relation avec un objet textuel n'appartenant pas à la SE.

Cette typologie a permis de définir ce que Luc et Virbel appellent les énumérations parallèles et non-parallèles. Celles-ci peuvent être définies comme suit :

- les énumérations parallèles sont *paradigmatiques*, *homogènes* et *isolées*.
- les autres sont appelées non-parallèles.

Exemples de SE Nous donnons ci-dessous deux exemples analysés selon la typologie de Luc. Pour chacun des exemples, nous donnons un commentaire, son graphe architectural ainsi que l'arbre RST correspondant.

L'exemple (3.c), issu de (Kamel et Rothenburger, 2011), montre une SE où la coénumérabilité est assurée par les dimensions syntaxique, rhétorique et visuelle.

Syntaxiquement, tous les items présentent un syntagme nominal avec une tête lexicale identique et complètent la phrase introduite dans l'amorce (qui est donc syntaxiquement incomplète[7]). Rhétoriquement, une même relation noyau-satellite lie l'amorce à l'énumération. Ainsi, l'énumération est dite *paradigmatique*. Visuellement, les marques typographiques (présence de ■) et dispositionnelles (retours à la ligne, retraits) sont identiques et renforcent la coordination entre les items. L'énumération est dite *homogène*. Comme l'énumération est également *isolée*, c'est-à-dire qu'elle n'entretient pas de relation avec des objets textuels externes, celle-ci peut être dite *parallèle*.

(3.c)

Les formes de communication non parlées sont :

- ■ le langage écrit
- ■ le langage des signes
- ■ le langage sifflé

Afin d'exemplifier les représentations MAT et RST, nous effectuons une segmentation en propositions de l'exemple (3.c) pour obtenir le résultat donné en (3.d). Les unités obtenues sont mises entre crochets et associées à un identifiant unique. Ces unités correspondent à la fois aux unités textuelles (UT) et à la fois aux unités discursives élémentaires (EDU). Il est important de rappeler que cela n'est pas systématique[8].

[7] Porhiel (2007) utilise le terme de non-saturé.

[8] Se référer à la section 2.1.3 ainsi qu'aux chapitres 6 et 7 de (Luc, 2000).

(3.d) [Les formes de communication non parlées sont :]$^{3.dA}$
- [le langage écrit]$^{3.dB}$
- [le langage des signes]$^{3.dC}$
- [le langage sifflé]$^{3.dD}$

En figure 3.1, nous donnons le graphe architectural correspondant à l'exemple (3.c). Les unités textuelles (UT) correspondent respectivement aux propositions issues de la segmentation (UT1 pour $3.dA$, UT2 pour $3.dB$, etc.). Les arcs en pointillé correspondent à la métaphrase de *composition* dans le métadiscours. L'objet textuel *énumération* est *chapeauté* par l'objet textuel *amorce*. Cette métaphrase est représentée par un arc de ligne continue avec une double flèche. Les items sont également représentés comme des objets textuels. Ceux-ci *agencent* l'énumération. Cette métaphrase est représentée par un arc à ligne continue avec une seule flèche. Il est important de noter que l'énumération n'est pas composée ici. Ceci permet de représenter des structures entrelacées.

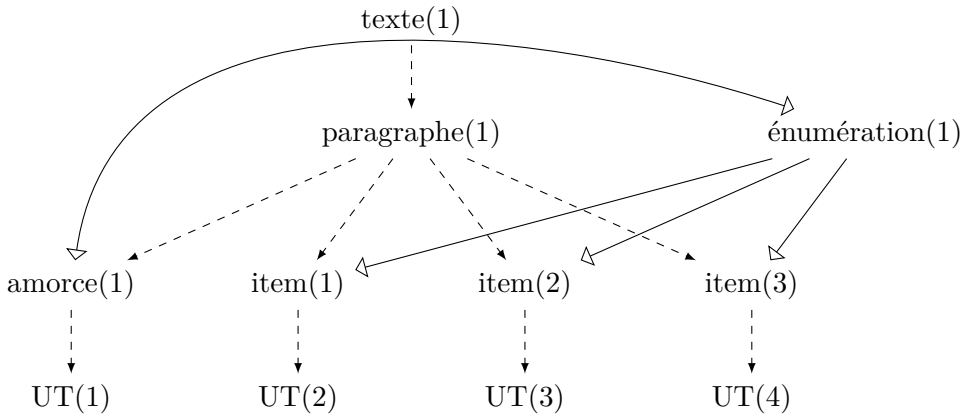

FIGURE 3.1 : Graphe architectural correspondant à l'exemple (3.d)

En figure 3.2, nous donnons l'arbre RST correspondant à l'exemple (3.c). Les propositions issues de la segmentation sont directement manipulées ici. L'amorce et l'énumération sont reliées par une relation noyau-satellite d'ÉLABORATION. Il s'agit d'une relation asymétrique qui est représentée ici par un arc plein partant du satellite ($3.dB - 3.dC - 3.dD$) pour atteindre le noyau ($3.dA$). Au sein de l'énumération, les items sont reliés par une relation multi-nucléaire de SÉQUENCE.

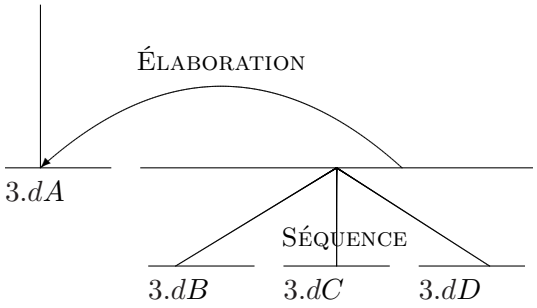

Figure 3.2 : Arbre RST correspondant à l'exemple (3.d)

L'exemple (3.e), issu de Virbel (1999)[9] et segmenté en propositions, montre une SE où les dimensions syntaxique, rhétorique et visuelle amènent des constructions distinctes.

Syntaxiquement, les trois premiers items présentent une structure verbale relativement identique, tandis que le dernier item paraît différent. Rhétoriquement, la structure est plusieurs fois imbriquée. Le premier item introduit la coordination du second et du troisième. Ceci est marqué par des répétitions de type lexical ainsi que par la présence de la conjonction de coordination « and ». Le quatrième item est subordonné au troisième par la conjonction de subordination « where ». L'énumération est dite *hybride*.

Visuellement, les quatre items sont équivalents. Les marques typographiques telles que la numérotation en début d'item, la ponctuation « ; » en fin des trois premiers items ainsi que les marques dispositionnelles (les retours à ligne, l'interligne, le retrait) assurent la coénumérabilité de l'ensemble de la structure. L'énumération est donc dite *homogène*. L'énumération est correctement *isolée*, cependant comme elle n'est pas *paradigmatique*, elle sera dite *non-parallèle*.

(3.e)

[In this paper I will defend what I shall call '(nonsolipsistic) conceptual role semantics'.]$^{3.eA}$ [This approach involves the following four claims:]$^{3.eB}$

(1) [The meanings of linguistic expressions are determined by the contents of the concepts and thoughts they can be used to express ;]$^{3.eC}$

(2) [the contents of thoughts are determined by their construction out of concepts ; and]$^{3.eD}$

(3) [the contents of concepts are determined by their 'functional role' in a person's psychology, where]$^{3.eE}$

(4) [functional role is conceived nonsolipsistically as involving relations to things in the world, including things in the past and future.]$^{3.eF}$

[9] Cet exemple a aussi été commenté par Luc. Notons qu'il est possible de trouver une version en ligne de cet exemple : http://www.nyu.edu/gsas/dept/philo/courses/concepts/NonSolips.html.

En figure 3.3, nous donnons le graphe architectural correspondant. Notons que tous les items sont mis au même niveau sur la base du critère de coénumérabilité. En figure 3.4, nous donnons l'arbre RST. Contrairement à la représentation rhétorique des énumérations parallèles, cette représentation est profondément imbriquée et reflète les relations induites par le contenu propositionnel et les connecteurs.

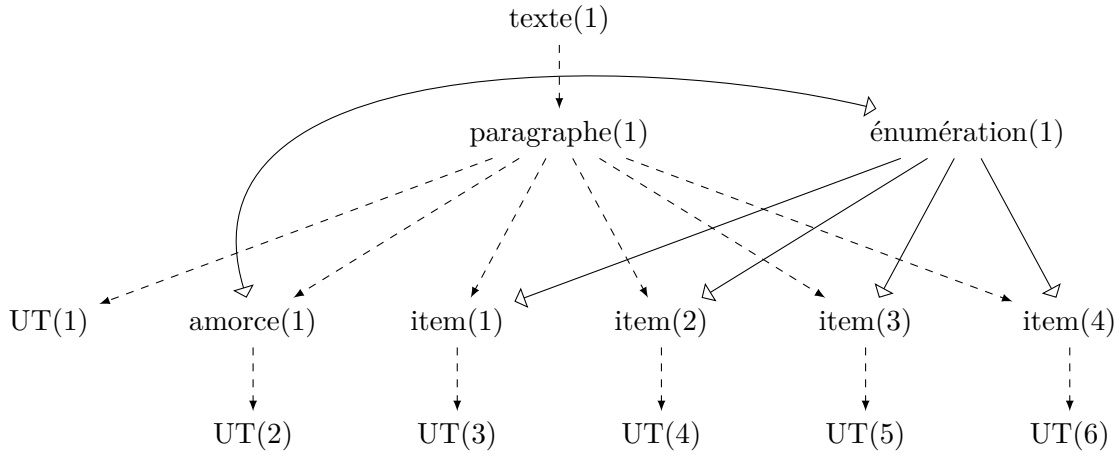

FIGURE 3.3 : Graphe architectural correspondant à l'exemple (3.e)

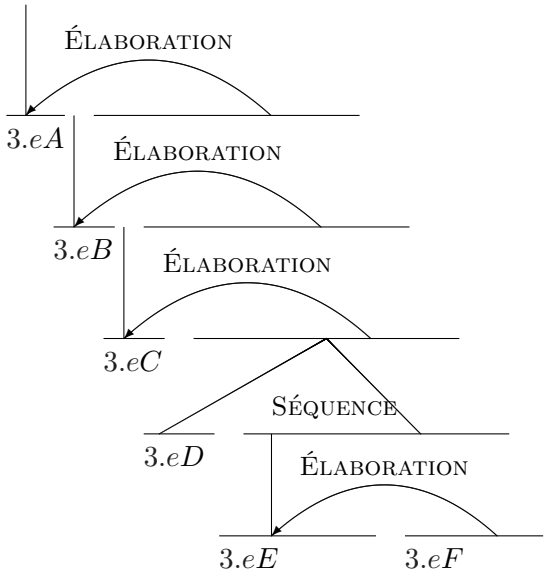

FIGURE 3.4 : Arbre RST correspondant à l'exemple (3.e)

3.2.2 Typologie de Ho-Dac, Péry-Woodley et Tanguy (2010)

Dans cette section, nous présentons la typologie de Ho-Dac, Péry-Woodley et Tanguy (2010). Celle-ci s'inscrit dans le cadre du projet ANNODIS, dont elle est un des résultats. Ici la SE est avant tout considérée comme une structure identifiable, à la manière de Turco et Coltier (1988)[10]. Dans cette section, nous exposons :

- le projet ANNODIS dédié aux SE ;
- la typologie résultante.

Étude des SE dans le cadre du projet ANNODIS Le projet ANNODIS[11] (Péry-Woodley *et al.*, 2009; Afantenos *et al.*, 2010) est un projet ANR démarré en 2007 dont l'objectif visait la construction d'un corpus annoté discursivement. Dans ce contexte, l'annotation du projet ANNODIS a exploré deux perspectives :

- une perspective ascendante qui démarre avec les unités discursives élémentaires et qui vise à mettre au jour des structures plus complexes au travers de l'annotation de relations du discours.
- une perspective descendante qui part du texte dans son ensemble et qui vise à identifier des structures de plus haut niveau, perceptibles de dans leurs dimensions visuelle et syntaxique.

La perspective descendante est celle dont nous discutons ici. Il s'agit de partir du texte d'un point de vue global en adoptant des méthodes de linguistiques de corpus et en utilisant des techniques TAL pour le pré-marquage des textes. Deux structures multi-échelles sont étudiées : les structures énumératives et les chaînes topicales.

Pour les SE, l'objectif poursuivi était d'étudier les interactions entre la structure logique du document et les indices lexico-syntaxiques (p. ex. la présence de lexèmes déictiques tels que *ci-dessous, précédemment*, etc.). Dans ce contexte, un intérêt particulier est donné à la signalisation et au passage entre indices et marqueurs. Ici, la signalisation est pensée aux travers de faisceaux d'indices à la fois lexicaux, syntaxiques, visuels (Péry-Woodley, 2000). Pour cela, le pré-marquage s'appuie à la fois sur des analyses morphologique et syntaxique et à la fois sur des « patrons ponctuationnels et typo-dispositionnels » (Ho-Dac *et al.*, 2010). Ceci permet de sortir l'annotateur d'une lecture purement linéaire.

L'annotation a été réalisé sur des textes expositifs de différents types (articles de Wikipédia, articles du CMLF[12] et rapports de l'IFRI[13]). La tâche d'annotation en elle-même a nécessité que l'annotateur délimite les SE et leurs composants constitutifs (amorce, item, etc.). Le corpus issu de cette annotation est accessible librement[14].

[10] « Les structures discursives recherchées se caractérisant par leur capacité à être perçues avant l'interprétation du contenu propositionnel qu'elles organisent, le rôle de la signalisation à la surface du texte est primordial. » (Ho-Dac *et al.*, 2010)
[11] http://w3.erss.univ-tlse2.fr/annodis
[12] Congrès Mondial de Linguiste Française
[13] Institut Français des Relations Internationales
[14] http://redac.univ-tlse2.fr/corpus/annodis/annodis_me.html

Typologie des structures énumératives Au terme de l'annotation, Ho-Dac *et al.* (2010) ont montré notamment que (i) les SE sont très présentes dans les textes expositifs et avec une couverture relativement large[15], et (ii) que les SE peuvent apparaître à des grains très différents, allant de l'ensemble de propositions à l'expression hiérarchique d'un chapitre entier. C'est sur ce dernier point que les auteurs ont proposé une typologie. Quatre types de SE sont distingués :

Type 1

SE dont l'énumération couvre des sections titrées.

Type 2

SE dont l'énumération correspond à des listes formatées visuellement.

Type 3

SE dont l'énumération est multi-paragraphique sans marques visuelles.

Type 4

SE dont l'énumération est intra-paragraphique.

Cette typologie présente deux avantages. Premièrement, elle permet de scinder l'ensemble des SE annotées en catégories relativement équilibrées. Deuxièmement, ces catégories montrent des associations significatives avec d'autres caractéristiques des SE. Par exemple, il est notamment montré que les SE de Types 1 et 2 présentent un plus grand nombre d'items que les autres, ou que les SE de Type 2 sont marquées plus fréquemment par la présence d'un classifieur (appelé ici énuméraThème).

3.3 Analyse sémantique des structures énumératives

Bien que les SE puissent montrer des discontinuités entre leurs composants, elles présentent un tout du point de vue sémantique. Exploiter cette partie sémantique est un travail qui a déjà été proposé sous certaines formes dans la littérature.

Usuellement, ces approches exploitent ce que Porhiel dénomme la « structure énumérative prototypique ». Celle-ci présente un classifieur qui est « lexicalement et sémantiquement réalisé par les co-items » (Porhiel, 2007). Dans ce cas, les items saturent ce classifieur et peuvent êre catégorisés avec un type équivalent à celui du classifieur :

> En ce qui concerne le classifieur, il « sert à définir la nature des items de l'énumération » (Maurel *et al.*, 2002), « la relation annonce/item [étant] toujours implicitement de type catégoriel » (Honeste et Froissart, 2003, p. 264). Ceci correspond à ce que l'on appelle une énumération classique, c'est-à-dire une énumération exprimant une relation sémantique d'hyperonymie.

[15] Dans les chiffres reportés par Hodac *et al.* (2010), il possible de constater que sur 56 documents, les annotateurs ont identifié 708 SE, soit une moyenne de 12,6 SE par documents. En outre, les auteurs montrent également que, en moyenne, 46% du contenu textuel est compris dans au moins une SE.

Notons également le travail de Gala (2003) dans lequel sont distingués deux types de relations entre le classifieur et les items d'une SE[16] : l'hyperonymie et l'holonymie. L'auteur conclut que le premier cas est plus fréquent que le second (Gala, 2003, p. 71) :

> La relation d'holonymie-méronymie est moins fréquente que la relation d'hyperonymie-hyponymie : sur cinquante listes extraites au hasard de corpus variés, après vérification manuelle, 36 listes (72%) sont des cas d'hyperonymie-hyponymie, 3 listes (6%) des cas d'holonymie-méronymie, le reste étant de cas n'appartenant à aucune de ces deux possibilités.

Ce propos est rejoint par celui de Aït-Mokhtar *et al.* (2003) qui suggèrent une correspondance entre les dépendances syntaxiques et les dépendances sémantiques qui lient les items à l'amorce[17].

Dans ce chapitre, nous évoquons quelques travaux exploitant ce plan sémantique. Nous divisons le chapitre en distinguant les approches qui exploitent des SE horizontales de celles qui exploitent les SE verticales.

3.3.1 Exploitation des structures énumératives horizontales

L'exploitation des SE horizontales a notamment été proposée dans le cadre de l'extraction de relations et de l'acquisition d'axiomes pour la construction d'ontologies. Nous présentons dans la suite ces cadres avec les travaux associés.

Extraction de relations L'exploitation des SE horizontales en extraction de relations a été proposée par Hearst (1992) et, à sa suite, Morin (1999). Les travaux de ces deux auteurs ont déjà été discutés en section 1.2.1 et inscrits au sein des approches contextuelles en extraction de relations. Nous présentons ci-dessous ces travaux en centrant notre propos spécifiquement sur l'exploitation des SE.

Hearst (1992) ne parle pas de SE, mais utilise le terme de liste L'intuition est que les termes qui occurrent dans des listes ont tendance à être liés sémantiquement[18]. Dans ce contexte, Hearst propose des patrons lexico-syntaxiques pour capturer les syntagmes nominaux énumérés. Par exemple, le patron ci-dessous est proposé :

$$such\ NP\ as\ \{\text{NP, }\}^*\ \{(or|and)\}\ NP$$

[16] Notons que Gala fait la distinction entre les *énumérations*, qui correspondent aux SE horizontales, et les *listes*, qui correspondent aux SE verticales (Gala, 2003, p. 65).

[17] « When the indicator phrase is a syntactic argument of a head in the introduction (...), the presence of the placeholder gives a good indication to extract a (semantic) relation between the list items and the introduction. In other words, for these particular cases there seems to be quite a direct "mapping" between a syntactic dependency (the one between the indicator and the head of the introduction) and a semantic dependency (the one between each list item and the introduction). » (Aït-Mokhtar *et al.*, 2003)

[18] « We observe that terms that occur in a list are often related semantically, whether they occur in a hyponymy relation or not. » (Hearst, 1992).

Celui-ci permet de traiter des exemples tels que donné en 3.f et d'en extraire des relations telles que `Hyperonymie("author","Herrick")`[19].

(3.f) (...) works by such authors as Herrick, Goldsmith, and Shakespeare.

Morin (1999) propose de traiter des « successions de syntagmes nominaux » afin d'en extraire des relations. Pratiquement, Morin propose d'utiliser des marqueurs de début d'énumération (p. ex. *tels que*, *comme*, les deux points), de coordinations entre items (p. ex. lettre avec parenthèses, tiret) ainsi que de fin d'énumération (p. ex. la locution *et cetera*). Ces marqueurs interviennent dans un ensemble de règles qui permettent d'extraire les listes de syntagmes nominaux. Ensuite, à l'aide d'une méthode d'amorçage[20] comparable à celle de Hearst[21], il est possible d'acquérir de nouveaux patrons. Les expériences de ce travail montrent que lorsque les SE endossent un rôle d'exemplification, elles sont très productives en relations d'hyperonymie. Considérons l'exemple en (3.g).

(3.g) En outre, des organes tels que le foie, les reins, le poumon et le pancréas sont sous contrôle hormonal.

Il est possible alors d'extraire des relations telles que `Hyperonymie("organe","foie")`, etc. Notons que le traitement est uniquement fait sur des énumérations intra-phrastiques.

Acquisition d'axiomes Völker (2007) propose d'utiliser les énumérations, parmi un ensemble d'autres manifestations linguistiques, pour acquérir des axiomes de disjonction et améliorer le processus de construction d'ontologies.

Les ontologies exprimées au travers de logiques de description reposent sur l'hypothèse du monde ouvert. Cette hypothèse implique que si deux classes sont déclarées sans que l'une ne soit explicitement disjointe de l'autre, il est alors possible d'imaginer des instances appartenant à ces deux classes. Par exemple, il est possible d'imaginer une instance qui appartienne à la fois à la classe *chat* et à la classe *chien*. Pour contraindre la déclaration des classes, il est nécessaire d'introduire des axiomes de disjonction. Cependant, ceux-ci ne sont habituellement pas intégrés au sein des ontologies[22], ou alors sont, comme l'ont montré Rector *et al.* (2004), régulièrement mal utilisés.

Parmi les sources utilisées pour acquérir ces axiomes[23], Völker propose d'utiliser les énumérations. L'hypothèse est que la disjonction entre deux classes données est souvent

[19] Notons que Hearst ne fait pas la distinction entre la relation d'hyperonymie et la relation « d'instance ».
[20] Pour un état de l'art sur les méthodes d'amorçage, se référer à la section 1.2.3.
[21] « Le processus d'acquisition de schémas lexico-syntaxiques que nous avons conçu reprend et complète la méthodologie de Hearst (1992). » (Morin, 1999, p.62)
[22] Par exemple, l'ontologie DBpedia (`http://dbpedia.org`) ne compte que 20 axiomes de disjonction.
[23] Völker propose d'utiliser la structure des ontologies à étendre, les ressources textuelles qui leur sont associées ou encore des ressources lexicales telles que Wordnet.

directement reflétée dans le langage et, dans ce contexte, les énumérations peuvent être considérées comme énumérant des classes disjointes. Plus formellement, il s'agit pour une énumération donnée de syntagmes nominaux $NP_1, NP_2, \ldots, (and|or)\ NP_n$ d'extraire les concepts $c_1, c_2, \ldots, c_k$ dénotés par ces syntagmes nominaux et à considérer ces concepts comme mutuellement exclusifs[24]. Notons ici que la problématique de l'interprétation d'un syntagme nominal et son lien à un concept n'est pas soulevé par Völker[25].

Par exemple, considérons l'énumération donnée en (3.h)[26]. Dans ce cas, il s'agit de distinguer *pigs*, *cows*, *horses*, *ducks*, *hens* et *dogs* comme des classes disjointes.

(3.h)	The pigs, cows, horses, ducks, hens and dogs all assemble in the big barn, thinking that they are going to be told about a dream that Old Major had the previous night.

Notons ici qu'il s'agit avant tout d'un travail sur l'énumération, et non la structure énumérative dont l'amorce est ici ignorée. Dans la pratique, le repérage de cette énumération se fait sur la base de patrons lexico-syntaxiques comparables à ceux de Hearst. Notons également que Völker propose un patron pour capturer explicitement la disjonction.

3.3.2 Exploitation des structures énumératives verticales

L'exploitation des SE verticales a notamment été proposée dans le cadre de l'extraction de relations, la reconnaissance d'entités nommées et les systèmes de question-réponse. Dans cette section, nous présentons ces cadres avec les travaux associés.

Extraction de relations Plusieurs approches en extraction de relations exploitant des structures textuelles pouvant être affiliées aux SE verticales ont déjà été introduites dans la section 1.3.2. La difficulté liée à la définition et à la délimitation des SE (section 3.1) fait qu'il est difficile de statuer sur le phénomène réellement étudié par ces approches. Quelle limite existe-t-il entre une structure hiérarchique exprimée par des titres et une SE ? Sans détailler à nouveau ces approches, nous établissons un continuum entre celles qui ne parlent pas explicitement de SE ou d'énumérations et celles qui s'inscrivent explicitement dans la littérature relative aux SE.

Les travaux de Shinzato *et al.* (2004a), Sumida *et al.* (2008) et de Brunzel (2008) proposent de traiter des structures dont certaines peuvent s'apparenter à des SE, mais sans les définir. Pour les deux premiers travaux, l'intérêt se porte sur des structures textuelles

[24] Un score de confiance est additionnellement attribué.
[25] Nous renvoyons le lecteur à la section 1.1.2 discutant des nombreuses difficultés liées à l'interprétation des mots et des relations qui les lient.
[26] Exemple extrait de (Völker *et al.*, 2007).

qui se rapprochent des SE prototypiques de Porhiel (2007). Pour Brunzel, le regroupement de termes par chemin dans la structure du document se rapproche de l'hypothèse d'une grande SE globale[27]. Les travaux de Aussenac-Gilles et Kamel (2009), et dans leur suite Kergosien *et al.* (2010), sont conscients des phénomènes d'énumération, mais définissent la sémantique avant tout par les balises XML des documents qu'ils traitent. Enfin, plus explicitement, Laignelet *et al.* (2011) ainsi que Kamel et Rothenburger (2011) proposent de traiter des SE en s'inscrivant dans le cadre théorique de Luc et Virbel.

Reconnaissance d'entités nommées Le travail de Bush (2003) vise à extraire des entités nommées (EN) en utilisant des SE verticales. L'hypothèse est que l'amorce (appelée ici *déclencheur*) exprime le type des EN contenues dans les items (appelés ici *articles*). Extraire ce type facilite alors la classification des EN. Ce travail s'appuie sur le système d'acquisition d'entités nommées proposé précédemment (Jacquemin et Bush, 2000).

Bush considère que les amorces peuvent être décomposées en quatre composants sémantiques. Nous en donnons une courte définition ci-dessous :

- l'**introducteur** introduit l'énumération qui suit l'amorce et spécifie sa localisation spatiale dans le co-texte. Il est caractérisé par un lexème cataphorique.
- l'**organisateur** explicite l'organisation de l'énumération, c'est-à-dire l'organisation des items qui la composent (p. ex. liste, collection, etc.).
- le **classificateur** donne la nature des EN composant l'énumération. Morpho-syntaxiquement parlant, il s'agit d'un syntagme nominal au pluriel. Pour Bush, le classificateur donne le *genus*.
- le **noyau sémantique** est composé du classificateur et de ses modifieurs. Ces derniers apportent les *differentiae* du *genus*. Pour Bush, le noyau sémantique a un rôle définitoire.

Bien que certains de ces composants puissent être facultatifs, Bush explique que tous ces composants sont présents dans une amorce dite canonique. L'exemple (3.i)[28] donne une SE avec une amorce canonique.

(3.i)
The following is a list of universities with field camps.
- Georgia State University
- Ohio University
- University of Tennessee

Dans ce cas-ci, « *The following is* » est l'introducteur et localise l'énumération qui va suivre. « *a list of* » est organisateur qui spécifie la forme. Le noyau sémantique est « *universities with field camps* », et au sein de celui-ci le classificateur est « *universities* ».

Il est important de noter que l'analyse proposée par Bush est faite ici uniquement sur des SE dont les items sont des EN. En effet, son corpus d'étude a été constitué par des

[27] Cette hypothèse a été évoquée par Ho-Dac (2007).
[28] Repris de Bush (2003).

pages Web contenant certains motifs (p. ex. *list of*). Dans ce contexte, il est commun de trouver des SE avec des énumérations paradigmatiques.

Système de question-réponse Dans son travail, Falco (2014) propose d'utiliser les SE[29] et les tables pour améliorer les résultats des systèmes question-réponse pour les questions à réponses multiples. En particulier, l'hypothèse est faite que les réponses aux questions-listes soient formulées sous forme de SE. Considérons l'exemple suivant[30], pour répondre à la question :

Quelles sont les 8 étapes pour la fabrication de la chaux ?

Falco propose que le système soit capable de trouver et traiter le passage-réponse suivant :

(3.j)

Pour l'essentiel, les procédés de la chaux passent par les étapes fondamentales suivantes, illustrées à la Figure 2.3 :
- extraction du calcaire,
- stockage et préparation du calcaire,
- stockage et préparation des combustibles,
- cuisson du calcaire,
- broyage de la chaux vive,
- hydratation et extinction de la chaux vive,
- stockage,
- manutention et transport.

Dans la pratique, un travail de normalisation des documents HTML est effectué. Durant ce traitement, les SE sont linéarisées et leurs composants sont indexés. L'analyse de la question donne le type attendu. Ceci permet de rechercher les SE dont le classifieur correspond à ce type. L'extraction des réponses se fait sur la base de règles. Par exemple, si l'item débute par un verbe, tout l'item est sélectionné. Le système termine en faisant l'agrégation des réponses extraites à partir d'autres sources (plein texte et tables).

3.4 Discussion

Nous avons vu que les SE constituent un phénomène complexe dont la définition et la délimitation ne sont pas unanimement partagées. Cette richesse a amené un panel d'auteurs à travailler sur ces structures et à proposer des angles différents d'études. Plusieurs typologies ont été proposées en circonscrivant les dimensions considérées. Ce chapitre a également montré que certaines SE présentaient un intérêt particulier au sein du champ de l'extraction de connaissances. Dans ce cadre, plusieurs travaux ont exploité des SE (ou leur énumération) en le mentionnant explicitement ou non.

[29] Notons que Falco utilise des SE horizontales également.
[30] Exemple repris de (Falco, 2014, p.53).

De manière transversale, il apparaît que le choix du corpus a une incidence sur la forme des SE qu'il est possible d'y trouver et sur les propriétés qu'elles ont tendance à présenter. Par exemple, Bush (2003) a montré que les corpus Web présentaient de nombreuses SE[31], mais également que la mise en forme semble plus fréquemment suppléer les aspects purement lexicaux et syntaxiques. Sur ce dernier point, Bush parle de « réduction linguistique » qui montre « une véritable tendance à la réduction des mots dont l'absence ne pose pas de problèmes pour la compréhension » dans les pages Web. Ce constat se reflète également aussi dans les chiffres du corpus ANNODIS (Afantenos *et al.*, 2012), où le corpus tiré de Wikipédia présente un plus grand nombre de SE. Celles-ci sont également davantage marquées visuellement (prédominance des Types 1 et 2).

Il apparaît également qu'il est possible d'effectuer des recoupements entre la dimension rhétorique des SE et l'exploitation de leur plan sémantique. La section 3.3 a montré que les SE exploitées sémantiquement sont généralement celles dites *paradigmatiques* dans la typologie de Luc (2000), c'est-à-dire des SE où une relation de type noyau-satellite (généralement une élaboration) relie l'amorce à l'énumération, et où une relation multi-nucléaire relie les items. Dans ce cas, les SE ont tendance à porter une même relation sémantique entre l'amorce et chacun des items.

Ceci trouve écho dans les travaux d'Asher et Lascarides (2003), ainsi que ceux d'Adam (2012). Dans la SDRT (*Segmented Discourse Rhetorical Theory*), le lexique est considéré comme un indice important pour déterminer les relations rhétoriques entre les propositions (Asher et Lascarides, 2003). Cette hypothèse sera vérifiée empiriquement dans le travail d'Adam (2012). Dans celui-ci, l'auteur montre que les liens sémantiques marquant la relation d'élaboration peuvent être variés et une proposition est faite pour utiliser la cohésion lexicale comme un indice supplémentaire dans le parsing rhétorique.

Notre travail peut être considéré comme une démarche inverse : le fait qu'une SE soit paradigmatique constitue un indice quant à la présence d'une relation sémantique dans cette SE. Cette relation sémantique tient alors entre le classifieur de l'amorce et les entités textuelles (termes et entités nommées) introduites par les items. Ainsi, il est naturel d'envisager la SE comme une structure textuelle supplémentaire où identifier des relations sémantiques.

[31] « Il existe des énumérations dans tous les types de textes, mais elles sont particulièrement fréquentes dans les pages Web, parce que ces dernières exigent une structure claire qui facilite la compréhension du lecteur. » (Bush, 2003)

Deuxième partie

Modélisation et identification automatique de la structure de document

Chapitre 4

Modélisation de la structure de document

Sommaire

Dans ce chapitre, nous décrivons un modèle pour représenter la structure hiérarchique des documents. Ce modèle se positionne dans la suite des modèles théoriques proposés au sein de la communauté TAL (présentés en section 2.1) en proposant (i) une abstraction de la mise en forme, ainsi que (ii) une connexion forte avec la structure discursive.

Néanmoins, notre modèle se démarque en se positionnant dans une perspective d'analyse, et non de génération de textes. Une attention particulière a été donnée à l'efficience et à la simplicité du modèle afin de favoriser son implémentation. Ceci est permis en remplaçant une représentation en constituants, qui est habituellement employée, par une représentation en dépendances. Dans ce contexte, l'expressivité de la description est ici réduite aux phénomènes hiérarchiques.

Dans un premier temps, nous décrivons les réflexions ayant conduit à l'établissement de ce modèle : nous proposons une redéfinition des niveaux de structuration du document, et nous présentons une comparaison entre les représentations en constituants et en dépendances. Dans un second temps, nous définissons notre modèle et nous présentons les choix qui ont été faits. Ceux-ci concernent les types de dépendance et les étiquettes logiques utilisées. Dans un troisième temps, nous donnons une comparaison avec les autres modèles d'architecture du document proposés en TAL. Enfin, dans la discussion, nous revenons sur l'intérêt de ce modèle et présentons ses limites.

4.1 Redéfinition des niveaux de structuration du document

Bien que la majorité des travaux s'accorde sur le fait que plusieurs niveaux de structuration du document existent (visuel, logique et discursif), il n'existe pas de consensus quant aux frontières exactes entre ces niveaux (section 2.1.4). En centrant notre propos sur les éléments textuels, nous redéfinissons plus clairement ces structures et proposons d'affiner la structure logique.

- La **structure visuelle** d'un document est la forme dans laquelle celui apparaît. Les unités visuelles sont identifiées par des indices de nature typographique et dispositionnelle, qui peuvent suivre une convention liée au support (p. ex. papier A4, format numérique, etc.), au dispositif de production (p. ex. crayon, machine à écrire, outil de composition, etc.) ou au statut du document (p. ex. article de conférence, prospectus commercial, etc.). L'*unité élémentaire* est l'alinéa, c'est-à-dire un segment textuel encadré par deux moyens dispositionnels (p. ex. retours à la ligne, espaces, etc.)[1]. Plusieurs alinéas peuvent composer un bloc visuel, dit aussi *unité visuelle complexe*, lorsque l'espace les séparant est plus petit ou égal à l'interligne d'un document.

- La **structure logique** d'un document est définie comme un niveau abstrait ordonnant les *unités logiques élémentaires* et *unités logiques complexes* du document. Ces unités sont dites logiques, car elles participent à la compréhension du texte en endossant un rôle métadiscursif, c'est-à-dire indépendant de leur contenu propositionnel et informationnel. À ce niveau, nous redéfinissons deux sous-structures dont la distinction est graduelle :

  - La **structure logique de surface** d'un document est composée d'*unités logiques élémentaires*. Ces unités peuvent être un titre, un paragraphe, un item, une citation mais aussi l'alinéa. À ce niveau, l'étiquette de chacune de ces unités dénote le rôle métadiscursif (ou son absence pour l'alinéa) qu'elle joue dans le texte à un niveau local (p. ex. paragraphe, titre, etc.).
  - La **structure logique profonde** ordonne les unités logiques élémentaires de manière à former des *unités logiques complexes* et correspond à l'organisation du document telle que marquée par l'auteur. Les unités complexes peuvent être des structures hiérarchiques, des définitions, des théorèmes mathématiques, etc., et peuvent s'imbriquer, se chevaucher ou encore se superposer. Au sein de cette structure, un phénomène d'altération du rôle peut éventuellement apparaître. Une unité élémentaire considérée comme paragraphe lorsqu'elle est prise isolément peut porter une étiquette logique d'item au sein d'une structure hiérarchique à l'échelle de la page ou du document.

[1] Notons que cette définition de l'alinéa diffère de celle habituellement donnée, à savoir : « Séparation marquée par un blanc laissé au commencement d'un paragraphe, dont la première ligne est ainsi en retrait par rapport aux autres » (Trésor de la Langue Française).

- La **structure discursive** d'un document est la structure qui organise son *message*. Les *unités élémentaires* et *complexes* de discours sont liées les unes aux autres par des relations rhétoriques. Plusieurs typologies de ces relations ont été définies dans la littérature (Mann et Thompson, 1988; Asher, 1993).

Dans la pratique, il est difficile de distinguer nettement ces structures, car elles se superposent et entretiennent entre elles des liens d'interdépendance. Par exemple, un *bloc textuel* en début de document et présentant une fonte différente du reste du document peut, dans un premier temps, être étiqueté comme *paragraphe*, et, dans un second temps, être reconnu comme un *résumé* du message du document. Dans ce cas, la mise en forme visuelle et une convention partagée permettent au rédacteur de signaler et au lecteur de reconnaître un même rôle discursif.

La structure logique profonde est la structure abstraite que nous proposons de représenter par un arbre de dépendances. Ce choix nous démarque des autres modèles dont la représentation peut être réduite à un arbre de constituants. Afin de souligner les différences entre ces notions, nous proposons de revenir sur leurs différences historiques dans le domaine de l'analyse syntaxique.

4.2 Représentations en constituants et en dépendances

Les représentations en constituants et en dépendances sont des notions empruntées au domaine de l'analyse syntaxique. Dans cette section, nous exposons brièvement chacun de ces types d'analyses et, ensuite, nous présentons les différences entre leurs représentations respectives.

Analyse en constituants L'analyse en constituants est un type d'analyse syntaxique largement répandu en linguistique. Mel'čuk (1988) évoque trois raisons expliquant la prédominance de l'analyse en constituants sur l'analyse en dépendances. Premièrement, l'anglais a été l'une des premières langues étudiées, et non des langues plus flexibles quant à l'ordre de leurs mots telles que les langues slaves[2]. Cet intérêt a notamment été marqué au travers du structuralisme et du distributionnalisme américains. Citons les travaux de Bloomfield (1933), Wells (1947)[3] ou encore Harris (1961). Deuxièmement, la formalisation mathématique qui a été ajoutée par le courant générativiste au travers, notamment, des grammaires syntagmatiques (Chomsky, 1956), s'est très essentiellement destinée à une analyse en constituants. Chomsky l'explicite dans le séminal *Syntactic Structures*[4] (1957). Troisièmement, l'aspect central donné à la syntaxe dans la vision générativiste a mis de côté l'aspect sémantique davantage véhiculé par la vision en dépendances.

[2] Se réferer au travail de Dikovsky et Modina (2000) qui montre les méthodes d'analyse en dépendances utilisées en URSS dans les années 60 et 70.

[3] Wells propose le terme d'analyse en constituants immédiats (*immediate constituents*) (Wells, 1947).

[4] « Customarily, linguistic description on the syntactic level is formulated in termes of constituents analysis (parsing). » (Chomsky, 1957, p. 27)

Dans la figure 4.1, nous donnons un exemple d'analyse en constituants pour la phrase « Le jeune essaie un pull ». Les nœuds terminaux présentent les formes linguistiques, tandis que les nœuds non-terminaux présentent les catégories de celles-ci.

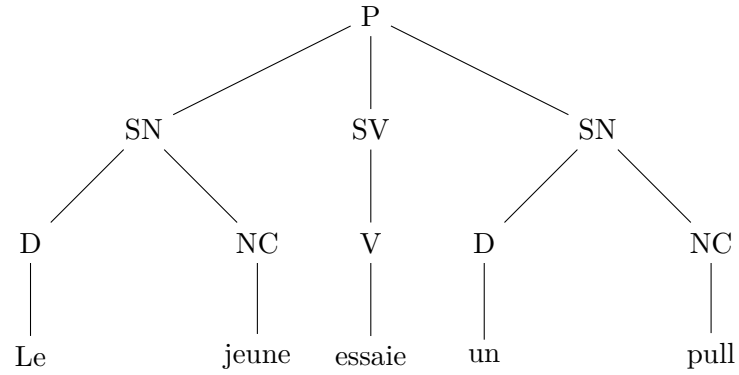

FIGURE 4.1 : Arbre de constituants pour la phrase « Le jeune essaie un pull »

Analyse en dépendances L'analyse en dépendances est un type d'analyse syntaxique dont l'usage tend à se répandre au travers d'approches computationnelles récentes (Nivre, 2008; Urieli, 2013). Historiquement, il est admis que le travail du français Tesnière (1959) constitue le socle fondateur de ce type d'analyse : le principe mis en avant est celui d'une dépendance entre une *tête* et son *dépendant*[5]. Il n'existe pas à proprement parler de grammaire de dépendances canonique, mais plutôt une multitude de cadres formels reposant sur le principe de dépendance (Kahane, 2000). Citons notamment la grammaire de Robinson (1970) ou la théorie Sens-Texte (Mel'čuk, 1988) (section 1.1.2).

La figure 4.2 donne un exemple d'analyse en dépendances pour la phrase « Le jeune essaie un pull ». Les dépendances sont typées avec les fonctions grammaticales.

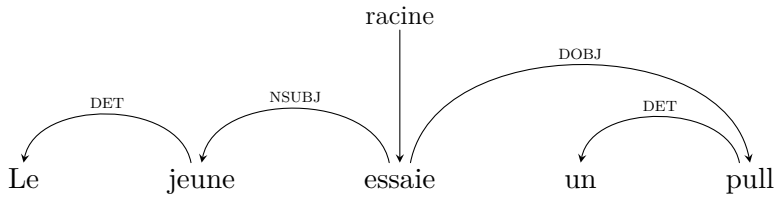

FIGURE 4.2 : Arbre de dépendances pour la phrase « Le jeune essaie un pull »

Notons que certains cadres formels utilisent des notions empruntées aux deux types d'analyses, telles que les Grammaires Catégorielles (Bar-Hillel *et al.*, 1960).

[5] Tesnière utilise les termes de régissant et subordonné : « 1. — Quand deux mots sont en connexion structurale, il y a **deux manières** de les placer en séquence linéaire, suivant que l'on commence par l'un ou par l'autre sur le relevé de la chaîne parlée.
2. — Dans un cas, on énonce **d'abord le régissant et ensuite le subordonné.** C'est ce qui se fait par exemple dans le français *cheval blanc* (v. St. 12).
3. — Dans l'autre cas, on énonce **d'abord le subordonné et ensuite le régissant.** C'est ce qui se fait par exemple dans l'anglais *white horse* (v. St. 13). » (Tesnière, 1959, p. 22)

Comparaison des représentations Pour la comparaison entre les arbre de constituants et les arbres de dépendances, nous reprenons les remarques faites par Mel'čuk (1988) dans sa défense de l'analyse en dépendances[6], mais également les réflexions de Kübler *et al.* (2009). Quatre points en particulier sont discutés :

composition et dépendance Le principe de composition de l'arbre de constituants évalue la manière dont les unités se combinent pour former des unités de plus haut niveau. La construction de l'arbre repose donc sur un lien de composition.

Le principe de dépendance met en avant la relation et son sens entre les unités plutôt que leur combinaison. L'intérêt est porté sur l'articulation régissant-subordonné, et la construction de l'arbre repose sur ces liens de dépendance.

catégorisation et fonction Les arbres de constituants encodent les combinaisons des unités au travers de catégories abstraites. En syntaxe, cette catégorisation est faite selon des catégories syntaxiques telles que le syntagme verbal (SV), le syntagme nominal (SN), etc. Dans ce cas, les catégories syntaxiques ne permettent pas de définir les fonctions grammaticales et celles-ci ne sont pas encodées dans l'arbre.

Les arbres de dépendances encodent des fonctions sur les arcs liant les unités. En syntaxe, il s'agit des fonctions grammaticales (nsubj, det, dobj, etc.). Dans ce cas, les fonctions grammaticales ne permettent pas de déterminer les catégories syntaxiques et celles-ci ne sont pas encodées dans l'arbre.

nœuds non-terminaux et terminaux Les catégories syntaxiques occupent les nœuds non-terminaux au sein des arbres de constituants, tandis que les nœuds terminaux sont liés aux formes linguistiques. Ceci implique que les arbres de constituants présentent plus de nœuds que les arbres de dépendances.

Les nœuds des arbres de dépendances trouvent une correspondance avec la forme linguistique analysée qu'ils soient terminaux ou non[7]. Les arbres de dépendances présentent donc moins de nœuds que les arbres de constituants.

ordre et flexibilité Les arbres de constituants nécessitent qu'un ordre soit donné aux nœuds liés aux formes linguistiques. Ceci implique que ce type d'analyse est plus apte à représenter des langues telles que l'anglais ou le français[8].

Les arbres de dépendances ne nécessitent pas d'ordre et permettent de représenter des langues où l'ordre des unités est plus flexible (p. ex. les langues slaves) au travers d'arbres non-projectifs.

[6] « This book has been written in order to plead the case for DEPENDENCY SYNTAX in modern linguistics. » (Mel'čuk, 1988)

[7] Cette correspondance n'est toutefois pas bijective. Il peut exister des nœuds vides ou, inversement, plusieurs nœuds peuvent représenter une même forme linguistique (Urieli, 2013).

[8] Néanmoins, cela n'est pas toujours vrai. Considérons l'exemple de McDonald *et al.* (2005b) *John saw a dog yesteray which was a Yorkshire Terrir.*

Les représentations des deux types peuvent être transformées de l'une à l'autre, mais cela n'est pas toujours immédiat. Un arbre de dépendances projectif peut être transformé en un arbre de constituants. Cela est montré en détail par Müller (2015). Également, un arbre de constituants peut induire un arbre de dépendances non typées (McDonald *et al.*, 2005b). La situation est plus complexe lorsqu'il s'agit d'obtenir les fonctions, qui ne peuvent pas être déduites directement par la forme de l'arbre (Candito *et al.*, 2009).

4.3 Modèle de représentation de la structure hiérarchique

Considérer l'analyse de la structure logique de manière équivalente à un problème d'analyse syntaxique a déjà été proposé (Section 2.2.2). Toutefois, employer explicitement l'analyse en dépendances et sa représentation pour traiter la structure logique des documents est, à notre connaissance, neuf et apporte au moins deux avantages :

(1) Reposer sur un principe de dépendance et non un principe de composition permet de mettre de côté la définition préalable de catégories abstraites occupant les nœuds non-terminaux. Ceci donne plus de flexibilité pour l'analyse et offre une alternative à la définition difficile de larges ensembles de contraintes de composition.

(2) L'arbre de dépendances offre une vue synthétique de la structure du document : l'articulation des nœuds reflète directement l'organisation du document, et tous les nœuds trouvent une correspondance immédiate avec les formes linguistiques. Ceci simplifie la visualisation et la manipulation de structures textuelles dans des traitements automatiques.

Dans la suite de cette section, nous donnons une définition formelle de notre modèle de représentation de la structure des documents, ensuite nous exposons les choix effectués. Ceux-ci concernent les types de dépendance et les étiquettes logiques utilisées. En fin de section, nous donnons un exemple d'analyse.

4.3.1 Définition formelle

La représentation de la structure logique profonde s'appuie sur un arbre de dépendances. Les relations de dépendance sont des relations binaires liant les unités logiques. Un nœud appelé *texte* et constituant la racine est ajouté au sommet de l'arbre. Nous décrivons ici cet arbre, ainsi que les contraintes syntaxiques que nous avons appliquées pour simplifier l'analyse logique des documents. Les travaux de formalisation de Nivre (2008), proposés dans un cadre d'analyse syntaxique, nous guident dans ces tâches.

Arbre de dépendances Pour un document donné, nous représentons son contenu par la séquence des unités logiques ul qui le composent et ordonnées selon l'ordre de lecture tel que :

$$d = (ul_1, ul_2, \ldots, ul_m)$$

Nous définissons l'ensemble T des types de dépendance et nous définissons le graphe de dépendances $G = (N, D)$ où :

- $N = \{n_0, n_1, n_2, \ldots, n_m\}$ est l'ensemble des nœuds du graphe où, excepté n_0, chaque nœud correspond à une unité logique tel que n_i correspond à ul_i.
- $D \subseteq \{(u, t, v) : u \in N, t \in T, v \in N\}$ est l'ensemble des arcs dirigés où chaque arc correspond à une dépendance, représentée par le triplet (u, t, v) où une tête u est liée à un dépendant v par la dépendance de type t.

Pour correspondre à un arbre de dépendances, le graphe G doit répondre à quatre contraintes syntaxiques :

(1) Le nœud n_0 est un nœud factice et ne peut être le dépendant d'un autre nœud.

(2) Pour un nœud donné u autre que n_0, il ne peut exister au maximum qu'une seule tête et un seul type de dépendance.

(3) Le graphe G est acyclique, c'est-à-dire qu'il n'existe pas de tête qui soit dépendante, immédiatement ou non, d'elle-même.

(4) Le graphe G est connexe, c'est-à-dire que si on remplace ses arcs dirigés par des arcs non-dirigés, le graphe est connecté.

Dans ce contexte, nous obtenons un arbre de dépendances possédant $|N|$ nœuds liés par $|N| - 1$ dépendances.

Contraintes Pour faciliter l'analyse logique du document, nous ajoutons deux contraintes syntaxiques à l'arbre de dépendances :

(1) **Projectivité des arcs** Un graphe de dépendances G est dit projectif si l'ensemble des nœuds atteignables par ses arcs le sont dans une fermeture réflexive et transitive. Autrement dit, pour deux unités logiques ul_i et ul_j, les unités logiques comprises entre i et j doivent être liées entre elles ou liées à ul_i et ul_j. Notons que cette contrainte de projectivité est une notion classiquement discutée en analyse syntaxique (McDonald *et al.*, 2005b; Candito *et al.*, 2009).

(2) **Transitions à droite** Dans la séquence d'unités logiques du document en ordre de lecture, seules les transitions à droite sont admises, c'est-à-dire que pour tout arc $(ul_i, t, ul_j) \in D$, i doit être strictement inférieur à j. Autrement dit, les arcs ne peuvent être orientés que vers les unités logiques entrantes durant le parsing.

Dans le cadre de l'analyse de documents, assurer la projectivité de l'arbre de dépendances et ne considérer que les transitions à droite facilite le processus d'analyse : l'attachement d'une nouvelle unité entrante ne peut se faire qu'à un nombre limité d'unités déjà traitées.

Dans la figure 4.3 nous donnons un exemple d'arbre de dépendances projectif et dont les arcs sont uniquement des transitions à droite. En figure 4.4, nous donnons un exemple d'arbre qui ne respecte pas les deux contraintes précitées.

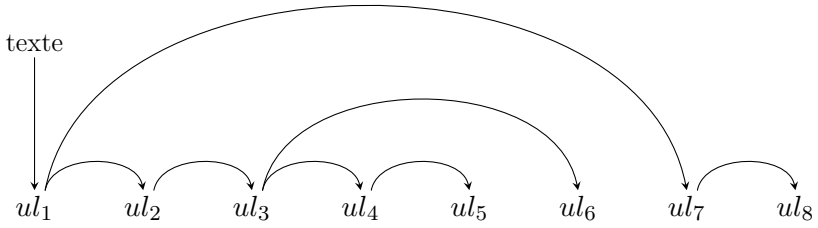

FIGURE 4.3 : Arbre de dépendances projectif avec transitions à droite

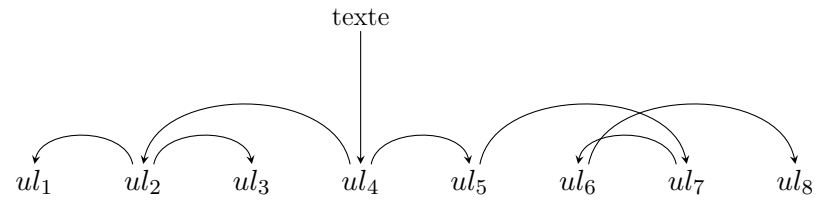

FIGURE 4.4 : Arbre de dépendances non-projectif avec transitions à gauche et à droite

4.3.2 Choix des types de dépendance

Dans notre modèle, deux types de dépendance sont considérés : la relation de subordination et la relation de coordination. Ces deux types de dépendance trouvent leur origine au sein des théories du discours. Dans cette section, nous montrons le parallèle avec ces cadres théoriques et, ensuite, nous présentons le principe de dépendance suivi.

Théories du discours La distinction entre subordination et coordination est faite dans de nombreuses théories du discours (Mann et Thompson, 1988; Polanyi, 1988; Asher et Vieu, 2005). Généralement, il est admis que la subordination apparie des éléments de premier plan avec des éléments de second plan, tandis que la coordination lie des éléments même plan (coordination/juxtaposition). Les différences entre ces théories apparaissent au niveau de la définition de ces relations et des contraintes appliquées.

Notre modèle reprend cette distinction et l'applique à la structure du document. Dans ce cadre, une relation de subordination désigne une descente dans la structure du document, et une relation de coordination lie deux unités partageant un même niveau et une même étiquette logique. Ainsi, notre modèle se rapproche des modèles théoriques simplifiés de Choi (2002) et de Hernandez et Grau (2005), avec néanmoins la différence que nous travaillons sur des unités logiques marquées visuellement, et non des propositions.

Principe de dépendance Le principe de dépendance suivi consiste à articuler ensemble les unités logiques qui apparaissent liées dans la cohérence du document. Dans la pratique, il s'agit d'exploiter les indices typo-dispositionnel et lexicaux pour choisir le type de dépendance adéquat liant deux unités logiques données.

Un parallèle peut être fait entre les marqueurs de cohésion utilisés dans notre modèle et ceux de la Mise en Forme Matérielle de Virbel (1989) (section 2.1.3). Trois types de marqueurs peuvent être considérés : (i) les marqueurs dispositionnels tels que les retours à la ligne, les retraits, etc., (ii) les marqueurs typographiques tels que les puces, les numérotations, la ponctuation, etc., et (iii) les marqueurs lexicaux, comprenant notamment les marqueurs d'intégration linéaire. Ces trois formes peuvent être combinées.

Il est important de noter que les dépendances de subordination et de coordination que nous proposons sont sous-spécifiées : la sémantique réellement véhiculée par ces relations n'est pas discutée ici. Par exemple, déterminer sémantiquement le lien entre un titre à un paragraphe sort du cadre de ce travail. Citons sur ce point l'étude de Lüngen *et al.* (2010) qui aborde en partie cette problématique et utilise la structure du document pour améliorer le parsing rhétorique.

4.3.3 Choix des étiquettes logiques

La question de la définition d'un ensemble d'étiquettes pour représenter la structure logique de document est une question difficile. Cette problématique a été évoquée pour la tâche d'analyse logique au sein de la communauté d'Analyse du Document (section 2.2.2). Nous avons montré qu'il n'existait pas de consensus et que les ensembles d'étiquettes étaient adaptés en fonction des objectifs poursuivis.

Dans la table 4.1, nous donnons deux exemples de travaux effectués en analyse logique avec leur ensemble d'étiquettes correspondant. Nous prenons ces travaux comme point de départ pour notre réflexion.

Travaux	Ensemble d'étiquettes logiques
Tsujimoto et Asada (1992)	titre, résumé, sous-titre, paragraphe, en-tête, pied de page, numéro de page, légende
Rangoni et Belaïd (2006)	titre, auteur, email, localité, résumé, mots-clefs, catégories, introduction, paragraphe, section, sous-section, sous-sous-section, liste, énumération, flottant, conclusion, bibliographie, algorithme, copyright, remerciements, numéro de page

TABLE 4.1 : Deux exemples d'ensembles d'étiquettes utilisés dans des travaux en analyse de la structure logique

Critiques des ensembles d'étiquettes Nous formulons trois critiques à l'encontre des ensembles d'étiquettes utilisés dans le tableau 4.1 :

1. Certaines étiquettes tiennent davantage de la structure visuelle plutôt que de la structure logique. Par exemple, les étiquettes en-tête, pied de page, numéro de page, etc. sont liées au support physique et ne devraient pas apparaître dans la structure logique.

2. Certaines étiquettes tiennent davantage de la structure rhétorique (ou du *message*). Si nous faisons l'hypothèse que la structure rhétorique est indépendante du support, alors des étiquettes telles que résumé, remerciements, conclusion, etc. ne sont pas des étiquettes liées à l'objet document.

3. Certaines étiquettes sont propres à une tâche ou à un type de document donné telles que copyright, algorithme, etc.

Définition des étiquettes Deux positions sur la définition des étiquettes peuvent être considérées. D'un côté, il est possible de multiplier les étiquettes afin de représenter finement les objets relatifs à un document ou à un domaine particuliers, tels que les définitions, les exemples linguistiques, etc. Nous avons vu dans le chapitre 2 que cette position était notamment celle d'Adobe pour le *Tagged PDF*, mais que ceci amenait des problèmes de cohérence.

D'un autre côté, il est possible de décomposer le texte en un ensemble restreint et fini d'unités logiques. La construction de nouveaux objets (p. ex. définition, etc.) se fait alors par juxtaposition et articulation d'objets atomiques. C'est notamment la position choisie par le langage de balisage HTML.

Nous inscrivons notre travail dans cette seconde position en utilisant un jeu d'étiquettes proche de ceux des langages de balisage (section 2.3.1). Il en découle que les étiquettes relatives à un mécanisme matériel (p. ex. pied de page, etc.) ou à un processus communicatif (p. ex. introduction) sont exclus de la représentation de la structure abstraite du document que nous proposons.

Notons que, idéalement, il serait plus cohérent de ne pas utiliser d'étiquettes, mais plutôt de proposer des classes d'équivalences visuelles. Sous l'hypothèse que, pour un document et un auteur donnés, une même mise en forme amène un même rôle logique, alors il serait envisageable de travailler avec des groupes sous-spécifiés. Néanmoins, cela n'est pas toujours possible ou efficient dans la pratique et une définition d'étiquettes reste généralement nécessaire. Nous reviendrons sur ce point dans les perspectives.

4.3.4 Exemple d'analyses

Dans cette section, nous opposons une analyse en constituants à une analyse selon notre modèle en dépendances. Ceci permet de montrer que notre modèle encode directement certains phénomènes linguistiques que les modèles en constituants ne peuvent pas représenter.

Considérons l'extrait du document `ling_poibeau`[9] en figure 4.5. Nous mettons de côté la question du choix des étiquettes logiques en proposant l'alphabet d'étiquettes $\Sigma = \{titre, paragraphe, item\}$ et nous proposons de représenter l'extrait par la séquence d'unités logiques élémentaires étiquetées suivante :

$$d = (\mathrm{ul}_1(\mathrm{titre}), \mathrm{ul}_2(\mathrm{paragraphe}), \mathrm{ul}_3(\mathrm{paragraphe}), \mathrm{ul}_4(\mathrm{item}), \mathrm{ul}_5(\mathrm{paragraphe}), \mathrm{ul}_6(\mathrm{item}), \mathrm{ul}_7(\mathrm{paragraphe}))$$

6.1 Un rayonnement limité de la recherche française

La fin des années 1960 voit exploser le domaine de la linguistique, au-delà du structuralisme. Ruwet importe la grammaire générative ; Gross importe les grammaires formelles puis développe une approche originale du traitement des langues naturelles, sur une base d'inspiration harrissienne. Culioli développe sa propre école avec une forte dimension cognitive, Quemada s'intéresse à la linguistique quantitative, *etc*. C'est aussi le temps des grands projets, le plus emblématique étant le lancement du dictionnaire de la langue française et le lancement conjoint de l'Institut Nationale de Langue Française (INaLF) à Nancy.

On assiste donc à un double mouvement.

- D'une part l'importation de théories élaborées à l'étranger, le plus souvent aux Etats-Unis, rend moins originales les recherches menées en France. Les chercheurs français traiteront ces théories d'origine anglo-saxonne avec un point de vue original, à l'image du regard critique de Milner sur la grammaire générative (1989) ou de Gross travaillant sur la base d'une analyse de type harrissien (1975). Toutefois, le point de vue français a une influence limitée au-delà des frontières : la France n'est plus le pays moteur en matière d'innovation et de création en linguistique.

  On assiste donc au développement d'écoles françaises sur la base de théories étrangères, mais l'écosystème linguistique français a des interactions limitées avec le monde extérieur. Par exemple, Harris développe à partir des années 1960 sa théorie des sous-langages sur une base distributionnelle mais les recherches de Gross restent relativement hermétiques à ces développements. Les deux chercheurs mènent dès lors des voies séparées et l'influence de Gross restera limitée. Quemada développe de son côté l'analyse lexicographique à partir de comptages systématiques sur corpus, mais ses recherches se développent indépendamment du monde anglo-saxon (même si des représentants de l'école anglo-saxonne ont assisté au grand congrès fédérateur organisé à Besançon en 1961, cf. Léon, 2004).

- D'autre part, les projets et les théories propres développés en France ont une audience limitée[6]. Culioli développe sans doute la théorie la plus originale de l'époque mais il écrit peu ; de fait, la théorie culiolienne des opérations énonciatives, qui aurait sans doute pu se développer beaucoup plus largement, reste méconnue à l'étranger et limitée à l'intérieur de la communauté française. Ce n'est que plus tard que les principaux écrits de Culioli seront réellement accessibles et diffusés (Culioli, 1991) mais sans doute trop tard pour être réellement influents sur un plan international.

Le bilan est donc contrasté : alors que la France est en pointe dans les années 1960, les innovations sont principalement le fait d'auteurs anglo-saxons. De fait, la place de la France décline relativement au niveau

FIGURE 4.5 : Extrait du document ling_poibeau

[9] Issu du corpus ANNODIS (Péry-Woodley *et al.*, 2009). Ce corpus a été discuté en section 3.2.2.

Une analyse en constituants de cet exemple, telle que réalisée en figure 4.6, (i) nécessite de définir préalablement plusieurs catégories abstraites occupant les nœuds non-terminaux (en capitales dans l'arbre), et (ii) ne facilite pas l'identification des structures textuelles hiérarchiques (p. ex. la structure constituée de ul_3, ul_4, ul_5, ul_6). Additionnellement, cette représentation implique un plus grand nombre de nœuds comparé au nombre d'unités logiques à lier. Ce type de représentation s'apparente à ce qui pourrait se retrouver dans un langage de balisages tels que HTML ou LaTeX où l'inclusion des balises est nécessaire.

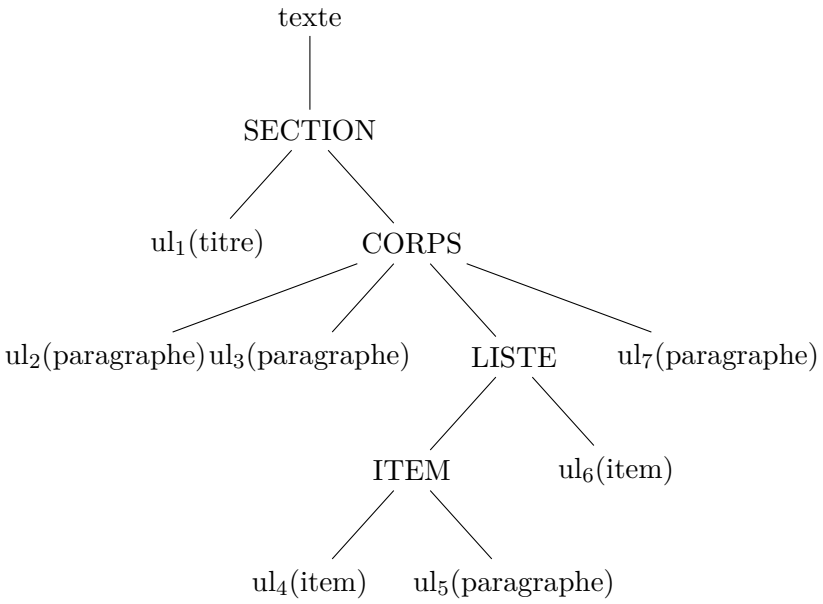

FIGURE 4.6 : Arbre de constituants pour l'exemple issu de ling_poibeau en figure 4.5. Les nœuds non-terminaux sont occupés par des catégories abstraites (en capitales). Les nœuds terminaux correspondent aux unités logiques élémentaires étiquetées.

Une analyse en dépendances du même exemple est donnée en figure 4.7. Dans ce cas, la définition de catégories abstraites n'est plus nécessaire, simplifiant dès lors le processus d'analyse. En outre, la représentation permet d'encoder directement la subordination entre le premier item (ul_4) et le paragraphe (ul_5) qui l'élabore, ainsi que la coordination (marquée visuellement et lexicalement) entre les deux items (ul_4 et ul_6). Ces phénomènes ne peuvent pas être directement encodés dans l'arbre de constituants. Notons que nous faisons ici une représentation verticale de l'arbre de dépendances. Ce qui est équivalent à la représentation horizontale faite en section 4.3.1.

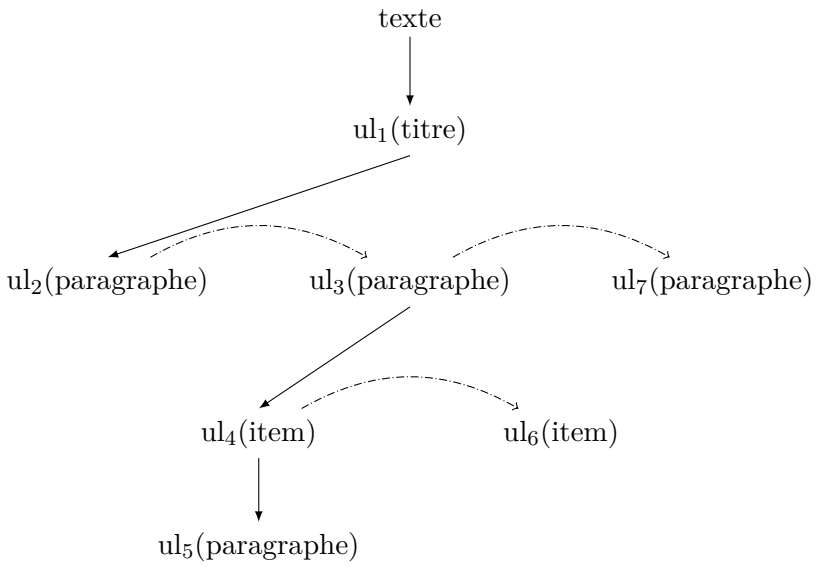

FIGURE 4.7 : Arbre de dépendances pour l'exemple issu de ling_poibeau en figure 4.5. La relation de subordination est représentée par une flèche pleine. La relation de coordination est représentée par une flèche en pointillé.

4.4 Comparaison avec les modèles théoriques en TAL

Nous donnons dans cette section une comparaison synthétique entre notre modèle de représentation de la structure logique et les autres modèles rendant compte de l'architecture des documents présentés dans le chapitre 2. Pour rappel, ces modèles étaient le modèle de Power *et al.* (2003), ci-après MDS, le modèle de Bateman *et al.* (2001), ci-après MBAT, et le modèle de Virbel (1989), ci-après MAT.

Notre comparaison porte sur deux axes : l'origine du modèle et la relation entre les unités logiques.

Origine du modèle Notre modèle a été conçu dans le but d'identifier des relations sémantiques à travers des structures hiérarchiques. Ce besoin nous a amenés à proposer un modèle permettant l'identification et la manipulation aisée des structures hiérarchiques marquées visuellement. Le MDS, le MBAT, le MAT trouvent leur origine dans le domaine de la génération de textes. Ceci implique que les représentations qu'ils proposent ne sont pas ou peu adaptables à une perspective d'analyse à partir de la structure visuelle. Dans ce contexte, les représentations qu'ils offrent sont soit très contraintes pour limiter le nombre de sorties générées, comme c'est le cas dans le MDS, soit très complexes comme c'est le cas dans le MBAT et le MAT qui manipulent des graphes et un grand nombre d'étiquettes. Notre modèle propose une représentation avec des contraintes adaptées pour une l'analyse automatique. Ainsi, il est naturel que notre modèle ne soit pas aussi fin que le MDS, aussi englobant que MBAT ou encore aussi expressif que le MAT.

Relation entre unités logiques Comme évoqué précédemment (Section 4.3), nous rejetons un principe de composition pour lui préférer un principe de dépendance. Ce choix va à l'encontre des autres modèles théoriques qui utilisent tous une relation de composition dans leur représentation. Nous montrons ici que cela implique une difficulté liée à la définition des catégories abstraites occupant les nœuds non-terminaux. Chacun de ces modèles propose une stratégie différente pour traiter cette difficulté :

- Dans le MDS, les auteurs contournent la difficulté en occupant les nœuds non-terminaux de l'arbre par des unités qui peuvent également occuper les positions terminales mais sans que leur correspondance avec la forme linguistique ne soit assurée[10]. Ceci permet la composition (au travers des règles de réécriture) en utilisant un nombre limité d'étiquettes, mais nécessite d'avoir des unités très fines.

- Dans le MBAT, les auteurs utilisent des conteneurs sans étiquettes et sans correspondance avec la forme linguistique. Notons qu'il peut y avoir des étiquettes, mais celles-ci sont arbitraires (p. ex. *rules* dans un document parlant de sport) et non systématiques[11]. Cette liberté pour l'étiquetage ainsi que l'absence de contrainte structurelle sur la représentation impliquent une représentation très riche, mais difficilement utilisable dans la pratique. Ce point est également souligné par Power *et al.* (2003)[12].

- Dans le MAT, les auteurs évitent la définition de catégories abstraites en permettant la composition récursive d'objets textuels. Toutefois, il ne paraît pas y avoir une liste finie d'objets textuels (ou du moins relativement limitée), et ceci implique l'établissement de larges tables de composition (Luc, 1998). Cette richesse couplée à la large liste de métaphrases offre le traitement de phénomènes complexes, mais rendent le MAT difficile à adapter à l'analyse automatique.

En remplaçant le principe de composition par celui de dépendance, notre modèle permet de faire en sorte que tous les nœuds de l'arbre (terminaux et non-terminaux) correspondent à des formes linguistiques. Ceci diminue l'expressivité du modèle, mais facilite l'analyse de la structure hiérarchique des documents (Section 4.3.4).

4.5 Discussion

Dans ce chapitre, nous avons proposé un modèle pour représenter la structure logique des documents. Ce modèle se positionne dans la suite des modèles théoriques proposés pour rendre compte de l'architecture textuelle : une abstraction de la mise en forme et

[10] Notons plus généralement que la grammaire de Nunberg ne considère pas de symboles non-terminaux (Section 2.1.1).

[11] Bateman *et al.* (2001) n'expliquent pas clairement ce point.

[12] « (...) Bateman and his colleagues do not provide a detailed account of the formation rules for layout structure, or of the constraints on the mapping between the RST tree and the layout structure. We are unsure, for example, whether "layout structure" would include such patterns as sections, paragraphs, and bulleted lists ». (Power *et al.*, 2003)

une connexion forte avec la structure rhétorique sont faites. Toutefois, notre modèle se démarque par une perspective d'analyse automatique des textes.

Cette perspective d'analyse a nécessité de remplacer le principe de composition, généralement utilisé, par un principe de dépendance. Ce changement amène deux avantages : (i) il n'est plus nécessaire de définir des étiquettes abstraites avec des règles complexes d'inclusion, et (ii) la représentation ainsi obtenue offre une vue synthétique de l'organisation des documents, facilitant ainsi l'identification et la manipulation de structures textuelles hiérarchiques.

Du point de vue de la granularité des phénomènes architecturaux pris en compte par notre modèle, nous pouvons distinguer deux niveaux :

- Au niveau micro, notre modèle se rapproche des travaux de Hernandez et Grau (2005), Jackiewicz (2005) et Couto *et al.* (2004) dans leur volonté de segmenter le document en parties de discours en utilisant les structures fines et les séries linéaires. Ces auteurs considèrent la proposition. Notre modèle se situe au-dessus en considérant le bloc textuel.

- Au niveau macro, notre modèle se rapproche du travail de Virbel (1989) dans sa volonté de représenter les phénomènes hiérarchiques à l'échelle du texte. Toutefois, notre modèle privilégie la simplification du processus d'analyse à l'expressivité de la représentation.

Ce dernier point amène un aspect limitatif de notre modèle. Les contraintes syntaxiques appliquées à notre représentation ne permettent de représenter que des documents de manière hiérarchique. Dès lors, les phénomènes non-hiérarchiques (p. ex. structures entrelacées), qu'il est possible de trouver dans les prospectus commerciaux, les pages de magazine, etc., sont hors de portée.

Du point de vue rhétorique, notre modèle pourrait être considéré comme un modèle discursif simplifié au sens de Choi (2002). Les deux types de dépendance que nous considérons, la subordination et la coordination, trouvent une traduction directe dans la majorité des théories de discours. Ainsi, une relative superposition entre la structure logique et la structure rhétorique est possible. Cette position diffère légèrement des celles prises par les modèles précédents. Toutefois, sur la question de la définition des étiquettes, nous avançons que celle-ci ne doit concerner que l'objet document en lui-même. Dès lors, nous pensons que des étiquettes logiques telles que résumé, introduction, conclusion, etc. doivent être évitées.

Chapitre 5

Identification automatique de la structure de document

Sommaire

Dans ce chapitre, nous décrivons le système implémentant le modèle proposé dans le chapitre 4. En entrée, le système prend un document et, en sortie, un arbre de dépendances représentant les relations de subordination et de coordination entre les unités logiques du document est attendu. La construction est effectuée de manière ascendante au travers de trois tâches :

1. **Segmentation en blocs textuels** : Les blocs de mot sont groupés les uns aux autres de manière à segmenter le document en blocs textuels. Cette segmentation s'effectue au travers d'un outil d'analyse géométrique fournissant des indices visuels en sortie. Pour un document donné, l'ensemble de blocs textuels obtenus en sortie est considéré comme la structure visuelle de ce document.

2. **Étiquetage des blocs textuels** : Chaque bloc textuel est étiqueté par une étiquette logique. Cet étiquetage se fait sur la base d'un faisceau d'indices visuels. Pour un document donné, la séquence d'étiquettes logiques obtenue en sortie est considérée comme la structure logique de surface de ce document.

3. **Construction de l'arbre de dépendances** : Les unités sont liées les unes aux autres par des relations de subordination et de coordination. Cette articulation s'effectue sur des critères visuels et lexicaux, mais également sur les étiquettes logiques précédemment étiquetées. Pour un document donné, l'arbre de dépendances en sortie est considéré comme la structure logique profonde de ce document.

La figure 5.1 associe chacune de ces tâches à un type d'analyse : la première tâche s'apparente à de l'analyse géométrique, tandis que les deux tâches suivantes correspondent à un processus d'analyse logique.

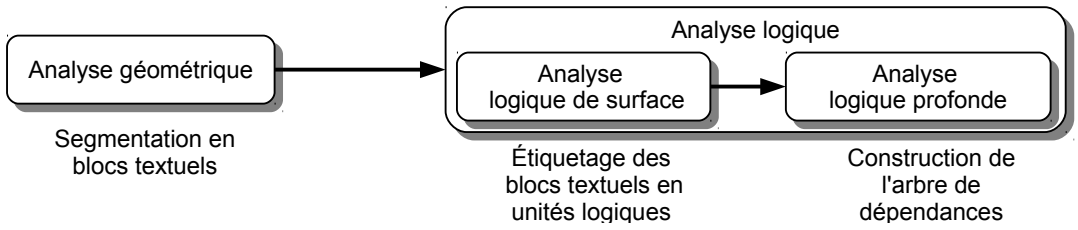

FIGURE 5.1 : Schéma du système pour l'identification automatique de la structure de document

Selon le type de format traité, la mise en œuvre de ces tâches est adaptée. Dans le cas des documents exprimés dans des langages de balisages, tels que le format HTML ou le format WikiText, la segmentation en blocs textuels et leur étiquetage sont déjà réalisés par les balises elles-mêmes. Seule la représentation du document sous la forme d'un arbre de dépendances est alors nécessaire. Le problème peut être résolu selon une approche déterministe avec des règles.

Dans le cas des documents exprimés par des langages de description de page, seule la structure visuelle est accessible. Ceci implique que les trois tâches sont nécessaires, et le problème doit être résolu en ayant recours à des modèles stochastiques.

Dans ce cadre, nous présentons la mise en œuvre des tâches pour le langage de description PDF, car celle-ci est nettement plus complexe que pour les formats à balises. Pour évaluer notre méthode, nous avons annoté un ensemble de documents PDF. Trois couches d'annotation correspondant respectivement aux structures visuelle, de surface et profonde ont été réalisées. L'annotation semi-manuelle de ces couches est décrite dans la première section de chapitre. Dans les sections qui suivent, nous décrivons les trois tâches et leur mise en œuvre pour les documents PDF.

5.1 Annotation semi-manuelle d'un corpus PDF

Nous avons choisi d'utiliser les corpus LING et GEOP issus du projet ANNODIS (précédemment décrit en section 3.2.2). Ces deux corpus présentent deux caractéristiques :

- une représentation originelle dans le langage de description PDF. Ceci nous offre un accès direct à la réalisation matérielle du document, contrairement aux documents exprimés en langage de balisage.
- l'annotation de la structure logique réalisée en amont de la campagne ANNODIS. Ces documents ont été transposés manuellement dans un format XML respectant la norme TEI[1]. Ceci ouvre la voie à une réutilisation du travail déjà effectué.

Le corpus LING est constitué de 25 articles scientifiques issus des actes du CMLF 2008[2]. Le corpus GEOP est constitué de 21[3] rapports/articles de l'IFRI[4]. La table 5.1 donne les caractéristiques générales de ces deux corpus.

	Documents	Pages
LING	25	308
GEOP	21	390
Total	46	698

TABLE 5.1 : Caractéristiques générales des corpus LING et GEOP

Nous avons enrichi semi-manuellement ces deux corpus par des annotations relatives à (1) leur structure visuelle, (2) leur structure logique de surface et, enfin, (3) leur structure logique profonde. Notons que les trois couches d'annotation sont librement accessibles[5] et modifiables conformément à la licence Creative Commons By-NC-SA 3.0[6].

5.1.1 Annotation de la structure visuelle

L'annotation de la structure visuelle vise à présenter chaque document PDF sous la forme d'une séquence de blocs textuels ordonnés selon l'ordre de lecture. Nous avons utilisé l'outil LA-PDFText[7], proposé par Ramakrishnan *et al.* (Ramakrishnan *et al.*, 2012), et avons corrigé manuellement les erreurs produites par cet outil. Notons qu'il existe des outils d'analyse géométrique autres que LA-PDFText fournissant des informations semblables et fonctionnant sur un principe relativement identique (p. ex. PDFX[8] (Constantin *et al.*, 2013)). Nous avons choisi LA-PDFText car il est distribué sous licence libre, ce qui nous a offert la possibilité d'affiner ses paramètres.

[1] Text Encoding Initiative - http://www.tei-c.org/
[2] Congrès Mondial de Linguistique Française
[3] Sur les 32 articles de GEOP, 21 ont été sélectionnés car considérés comme représentatifs des propriétés visuelles et logiques du corpus.
[4] Institut Français de Relations Internationales
[5] http://github.com/fauconnier/corpus-LING_GEOP
[6] https://creativecommons.org/licenses/by-nc-sa/3.0/
[7] Layout-Aware PDF Text Extraction - http://code.google.com/p/lapdftext/
[8] PDF-to-XML - http://pdfx.cs.man.ac.uk

Annotation automatique avec LA-PDFText L'outil LA-PDFText permet la segmentation automatique des pages de documents PDF en blocs textuels. Cette analyse géométrique repose sur une approche ascendante en trois étapes.

1. Dans un premier temps, les chaînes de caractères sont identifiées au travers de la librairie JPedal[9] dans sa version GPL. Ces unités atomiques sont appelées *blocs de mot* et sont caractérisées par leurs coordonnées (en pixels) sur la page, leur contenu textuel, leur fonte et la présence d'éventuelles marques d'emphases.

2. Dans un second temps, ces blocs de mot forment, par regroupement progressif, des *blocs textuels*. L'algorithme de regroupement utilisé dans LA-PDFText suit un principe glouton : pour un bloc de mot donné, les blocs de mot contigus sont unis s'ils apparaissent en deçà d'un seuil horizontal ou d'un seuil vertical calculés localement pour chaque page :
   - Le seuil horizontal est $\theta_{\text{horizontal}} = \text{plusCourant}(D_{horizontal})+\text{plusCourant}(H_{mots})$
   - Le seuil vertical est $\theta_{\text{vertical}} = \text{plusCourant}(D_{vertical}) + \text{plusCourant}(H_{mots})$

   où $D_{horizontal}$ et $D_{vertical}$ sont les distributions discrètes des distances horizontales et verticales entre blocs de mot, H_{mots} est la distribution discrète de la hauteur des blocs de mot et la fonction plusCourant(.) retourne la valeur considérée comme la plus courante dans la distribution donnée en argument.

3. Dans un troisième temps, une série d'heuristiques traitent les cas spécifiques en, par exemple, divisant verticalement les blocs textuels présentant des fontes différentes ou en joignant horizontalement les blocs textuels d'une même région qui partagent une même ordonnée.

En figure 5.2, nous donnons un exemple de regroupement des blocs de mot (ling_muller - corpus LING). Les parties grises sont les blocs textuels construits à l'état où l'algorithme évalue le bloc de mot 125. Le rectangle de bordure hachurée autour du bloc de mot numéro 125 est la zone calculées avec les seuils[10]. Les blocs des mot numéro 124 et numéro 126 forment une intersection avec cette zone. Notons que l'ordre des blocs de mot ne suit pas l'ordre de lecture. Ceci permet le traitement des mises en pages à multiples colonnes.

Correction manuelle des erreurs Deux types d'erreurs ont été observés. Le premier groupe d'erreurs concerne les problèmes de découpage. En fonction des seuils calculés sur chaque page (ou de l'application d'heuristiques au cas par cas), les résultats obtenus varient selon la convention graphique suivie. Par exemple, dans les documents du corpus GEOP, certains auteurs utilisent des espaces d'interligne larges. Ceci amène l'algorithme à considérer chaque ligne comme un bloc textuel à part entière. Le second groupe d'erreurs est l'absence de nombreux caractères de ponctuation. Lorsque ceux-ci ne sont pas

[9] Java PDF Extraction Decoding Access Library - https://www.idrsolutions.com/jpedal/

[10] Notons que, pour la clarté de cet exemple, nous avons divisé les seuils $\theta_{\text{horizontal}}$ et θ_{vertical} par deux. Dans la pratique, le regroupement est initialement fait avec une granularité moyenne, ensuite les heuristiques permettent une division plus fine.

3.1.2 L'inversion du sujet nominal

Il s'agit de diverses constructions à sujet nominal (non clitique) accordé au verbe qui le précède. Cette inversion a été nommée "inversion stylistique" (Kayne 1973). Il existe de nombreuses études depuis une trentaine d'années sur ce sujet. L'article de Kampers-Mahne, Marandin, Drijkoningen, Doetjes & Hulk (2004) distingue plusieurs types:

1. L'inversion dans les contextes d'extraction (questions partielles, relatives, clivées):

(43) Où est allée Marie? Je me demande où est allée Marie.

(44) La personne qu'a rencontrée Pierre est ma cousine.

(45) C'est dans cette maison qu'est né Victor Hugo.

2. L'inversion inaccusative (Marandin 2003), liée à des verbes spécifiques (verbes de mouvement, verbes avec auxiliaire *être*, passifs); elle est observable dans deux classes distinctes de contextes:

-complétives:

(46) Je voudrais que soient distribués ces prospectus (Kampers-Mahne et al., 2004: 557).

(47) On eût dit que traînait dans la pièce quelque chose de cette atmosphère lourde...(Gracq, Le rivage des syrtes, 32, Frantext).

-indépendantes avec ou sans adverbe introducteur:

(48) A ce moment-là se fit entendre un bruit strident.

(49) Entre alors notre gardien avec de la nourriture.

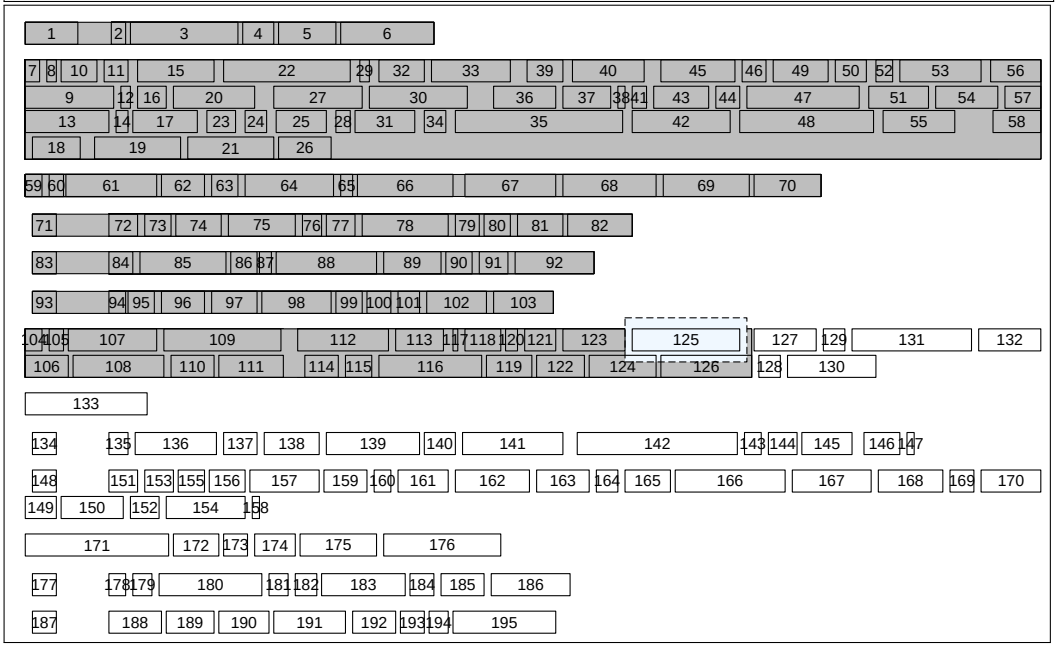

FIGURE 5.2 : Exemple de regroupement progressif des blocs de mot (en blanc) en blocs textuels (en gris) au moment de l'évaluation du bloc de mot numéro 125 dans un extrait du document ling_muller (corpus LING)

accolés à un mot, le système a tendance à ne pas les identifier. Ces erreurs peuvent s'expliquer par le traitement originel de l'anglais, où les règles typographiques impliquent que la ponctuation soit accolée aux mots.

Pour corriger ces deux types d'erreurs, nous avons inspecté l'ensemble des documents et avons manuellement rectifié les erreurs repérées.

Résultat La table 5.2 présente quelques caractéristiques des corpus LING et GEOP. LING présente des documents avec un nombre de pages restreint, mais un grand nombre de blocs textuels. La situation inverse apparaît pour GEOP.

Le format d'annotation est celui obtenu en sortie de LA-PDFText. Il s'agit d'un format XML avec des balises Page, Chunk et Word et une racine Document. Les balises présentent des attributs de nature dispositionnelle : les paires $(x_1,\ y_1)$ et $(x_2,\ y_2)$ représentent respectivement les coordonnées (en pixels) du coin supérieur gauche d'un bloc et de son coin inférieur droit. Additionnellement, les blocs de mot présentent des attributs de nature typographique. La figure 5.3 présente un extrait du corpus.

	Blocs de mot		Blocs textuels	
	Nombre	par page	Nombre	par page
LING	188 388	611,65	3839	12,46
GEOP	142 889	366,38	2676	6,86
Total	331 277	474,61	6515	9,33

Table 5.2 : Caractéristiques visuelles des corpus LING et GEOP

5.1.2 Annotation de la structure logique de surface

L'annotation de la structure logique de surface vise à représenter chaque document sous la forme d'une séquence de blocs textuels qualifiés avec des étiquettes logiques. De manière analogue à la phase d'annotation de la structure visuelle, cette phase fait intervenir une manipulation humaine. Nous avons défini un ensemble d'étiquettes logiques comparable à celui proposé dans les langages de balisage. Ensuite, nous avons récupéré les étiquettes déjà annotées dans le projet ANNODIS. Enfin, nous avons annoté manuellement les blocs textuels manquants avec une interface en ligne de commande.

Étiquettes logiques Deux difficultés peuvent être décrites ici. Premièrement, il est difficile de choisir un jeu d'étiquettes qui soit à la fois assez descriptif pour décrire les phénomènes présents sur une page et assez restreint que pour rester consistant au travers de plusieurs genres de documents. Notre position a été de restreindre le jeu d'étiquettes (Section 4.3.3). Deuxièmement, l'observation en corpus a montré qu'il n'existe pas d'association systématique entre la mise en forme visuelle et le rôle logique attendu d'un bloc textuel. Ceci nous a amenés dans cette phase d'annotation à respecter au maximum un principe d'une annotation *par la forme* et non par le *rôle logique* joué par un bloc textuel donné. Cette réduction permet de simplifier l'interprétation.

```
<Document>
  ...
 <Page x1="85" y1="142" x2="510" y2="699" type="page"
          chunkCount="23" pageNumber="7" wordCount="410">
   ...
  <Chunk x1="85" y1="317" x2="256" y2="327" type="unclassified" id="115">
   <Word x1="85" y1="317" x2="107" y2="327" font="Arial"
                  style="font-size:10pt;font-style:Bold">3.1.2</Word>
   <Word x1="121" y1="317" x2="127" y2="327" font="Arial"
                  style="font-size:10pt;font-style:Bold">L</Word>
   <Word x1="129" y1="317" x2="174" y2="327" font="Arial"
                  style="font-size:10pt;font-style:Bold">inversion</Word>
   <Word x1="176" y1="317" x2="189" y2="327" font="Arial"
                  style="font-size:10pt;font-style:Bold">du</Word>
   <Word x1="191" y1="317" x2="215" y2="327" font="Arial"
                  style="font-size:10pt;font-style:Bold">sujet</Word>
     ...
  </Chunk>
    ...
 </Page>
  ...
</Document>
```

FIGURE 5.3 : Format XML des propriétés visuelles d'un extrait du document ling_muller

Dans ce contexte, nous distinguons deux jeux d'étiquettes :

- Le premier jeu d'étiquettes correspond essentiellement à un sous-ensemble de ceux proposés dans les langages de balisage HTML et LaTeX. Les étiquettes choisies sont : les *titres* (Annexe A.2), les *paragraphes* (Annexe A.8), les *items* (Annexe A.7), les *citations* (Annexe A.10), les *titres de documents* (Annexe A.9), les *bylines*[11] (Annexe A.1) et les *références bibliographiques* (Annexe A.5). Ces étiquettes seront celles utilisées dans la dernière phase d'annotation visant la construction de l'arbre de dépendances.

- Le second jeu d'étiquettes répond à un besoin pratique : celui de pouvoir, dans une analyse du document, mettre de côté les blocs textuels qui ne sont pas utilisés dans la construction de l'arbre de dépendances. Nous considérons les étiquettes suivantes : les *en-têtes* (Annexe A.3), les *pieds de page* (Annexe A.3), les *notes de bas de page* (Annexe A.4), les *notes de fin* (Annexe A.6). Également, une étiquette *autres* a été été choisie pour les blocs textuels utilisés avec des éléments graphiques (p. ex. légendes d'images, contenus textuels des tables, etc.).

[11] Le terme *byline* est un terme générique utilisé pour désigner les alinéas généralement en début de document consacrés notamment à l'auteur, sa position et la date de parution, etc.

Notons que, dans notre version annotée librement accessible, une définition plus fine des étiquettes (p. ex. distinction entre titre de bibliographie et titre de niveau 1) est également proposée en vue d'éventuelles recherches complémentaires.

Annotation semi-manuelle L'annotation du corpus avec les étiquettes logiques proposées a été effectuée en deux étapes. Premièrement, une annotation a été réalisée avec l'algorithme de similarité textuelle de Myers (1986) en associant les étiquettes logiques du corpus ANNODIS originel aux blocs textuels de notre corpus sur la base du contenu textuel. Deuxièmement, à l'aide d'une interface en ligne de commande, les étiquettes non traitées dans ANNODIS (p. ex. *en-têtes*, *notes de bas de page*, etc.) ont été ajoutées manuellement.

Au terme de ces deux étapes, l'ensemble des blocs textuels issus de l'annotation de la structure visuelle ont été étiquetés et prennent l'appellation d'unité logique élémentaire. Par exemple, en figure 5.4, nous donnons l'annotation logique de surface pour l'extrait du document ling_muller.

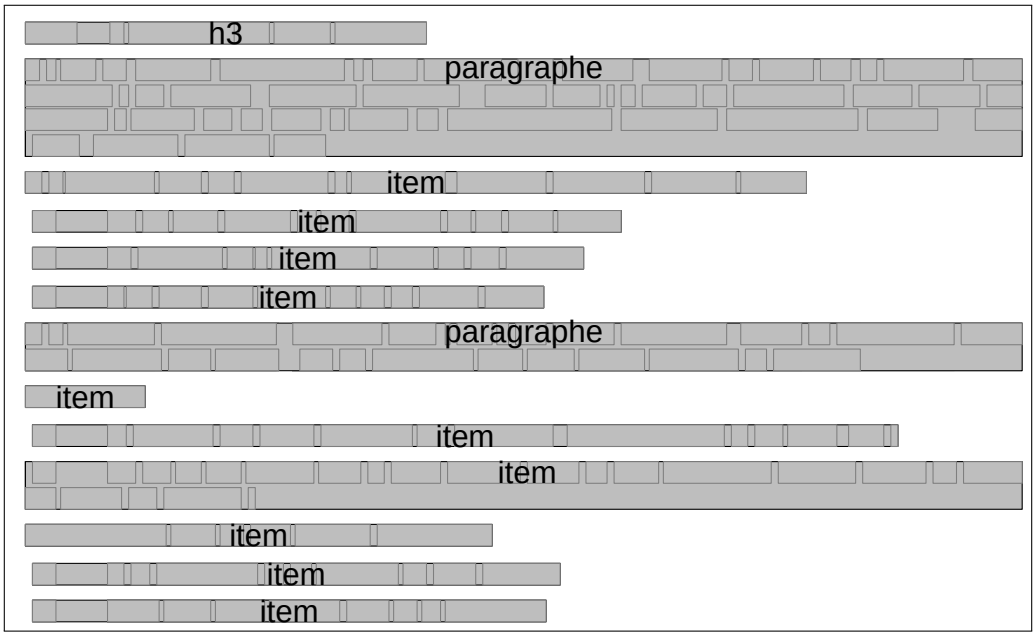

FIGURE 5.4 : Blocs textuels (en gris) avec leur étiquette logique respective dans un extrait du document ling_muller (corpus LING)

Résultats Des différences significatives (calculées par un χ^2 avec un risque α à 0,01) apparaissent dans les distributions des étiquettes de LING et GEOP (Table 5.3). Le caractère académique de LING implique un plus grand nombre d'items (dont 210 sont dédiés à l'énumération d'exemples linguistiques), ainsi qu'un grand nombre de citations et de références bibliographiques. La convention du CMLF implique que les notes de bas de page sont presque systématiquement remplacées par des notes de fin de document. Concernant le corpus GEOP, son caractère visuellement hétérogène s'observe au travers du grand nombre d'en-têtes et pieds de page, ainsi que par la présence de nombreux blocs textuels étiquetés *autres*. De manière transversale, le paragraphe est l'unité la plus représentée dans les deux corpus, légèrement en plus grand nombre dans GEOP dont les articles se veulent plus littéraires.

	LING	(25 doc.)	GEOP	21 (doc.)		
étiquettes	Nombre	Moyenne	Nombre	Moyenne	Total	Couv. %
titre	27	1,08	28	1,33	55	0,84
byline	53	2,12	96	4,57	149	2,29
réf. biblio.	1173	**46,92**	25	1,19	1198	18,39
h1	157	6,28	101	4,81	258	3,96
h2	108	4,32	78	3,71	186	2,85
h3	40	1,6	65	3,10	105	1,61
paragraphe	1241	49,64	1189	56,62	2430	**37,30**
item	380	**15,2**	72	3,43	452	6,94
citation	123	**4,92**	1	0,05	124	1,90
en-tête	45	1,8	171	**8,14**	216	3,32
pied de page	16	0,64	257	**12,24**	273	4,19
note de bas de page	7	0,28	370	**17,62**	377	5,79
note de fin	387	**15,48**	28	1,33	415	6,37
autres	82	3,28	195	**9,29**	277	4,25
Total	3839	153.56	2676	127.43	6515	100

TABLE 5.3 : Distribution des étiquettes logiques au sein de LING et GEOP

5.1.3 Annotation de la structure logique profonde

L'annotation de la structure logique profonde vise à présenter chaque document PDF sous la forme d'un arbre de dépendances typées ordonnant ses unités logiques élémentaires. Seules les unités participant au corps du texte sont considérées, c'est-à-dire la titraille, les paragraphes, les items, les citations, les bylines et les références bibliographiques.

Annotation semi-manuelle L'annotation de la structure logique profonde a été réalisée en trois étapes :

(1) Nous avons modifié manuellement les unités logiques élémentaires pour obtenir une représentation plane du document. Cette représentation plane écarte les étiquettes hors du corps du texte et regroupe les parties d'un même paragraphe coupé sur deux (ou plusieurs) pages (Annexe A.8). La notion de page disparaît également.

(2) Les arbres de dépendances ont été générés à partir d'une grammaire formelle décrivant avec des règles les dépendances entre les étiquettes logiques. Par exemple, ces règles statuent qu'un item est toujours subordonné au paragraphe qui le précède, ou encore que deux items sont systématiquement coordonnés. Ces règles ne sont naturellement pas toujours vérifiées. Nous reviendrons sur ces règles et la grammaire de dépendances qu'elles expriment ultérieurement[12].

(3) Nous avons corrigé manuellement les arbres de dépendances générés. Les erreurs concernaient essentiellement l'expression de structures textuelles imbriquées où l'indentation n'était pas respectée. Lorsque des cas ambigus apparaissaient, nous nous sommes référés à l'annotation de la structure logique qui a été faite dans ANNODIS pour trancher.

Au terme de ces étapes, les documents de LING et GEOP sont annotés avec des arbres de dépendances. En figure 5.5, nous donnons l'annotation correspondante à l'extrait du document ling_muller.

Résultats Une différence significative apparaît (χ^2 avec un risque α à 0,01) entre LING et GEOP : le corpus LING présente une structure beaucoup plus riche et profonde avec de nombreuses relations de subordination et de coordination (Table 5.4). Cette différence s'explique par les nombreux exemples linguistiques pour la relation de subordination, mais également les nombreuses références bibliographiques pour la relation de coordination.

	LING	(25 doc.)	GEOP	21 (doc.)		
étiquettes	Nombre	Moyenne	Nombre	Moyenne	Total	Couv. %
subordination	714	**28.56**	391	18.62	1105	24.02
coordination	2467	**98.68**	1029	49	3496	75.98
Total	3181	127.24	1420	67.62	4601	100

TABLE 5.4 : Distributions des dépendances typées au sein de LING et GEOP

[12] La description formelle de la grammaire est donnée en section 5.4.1.

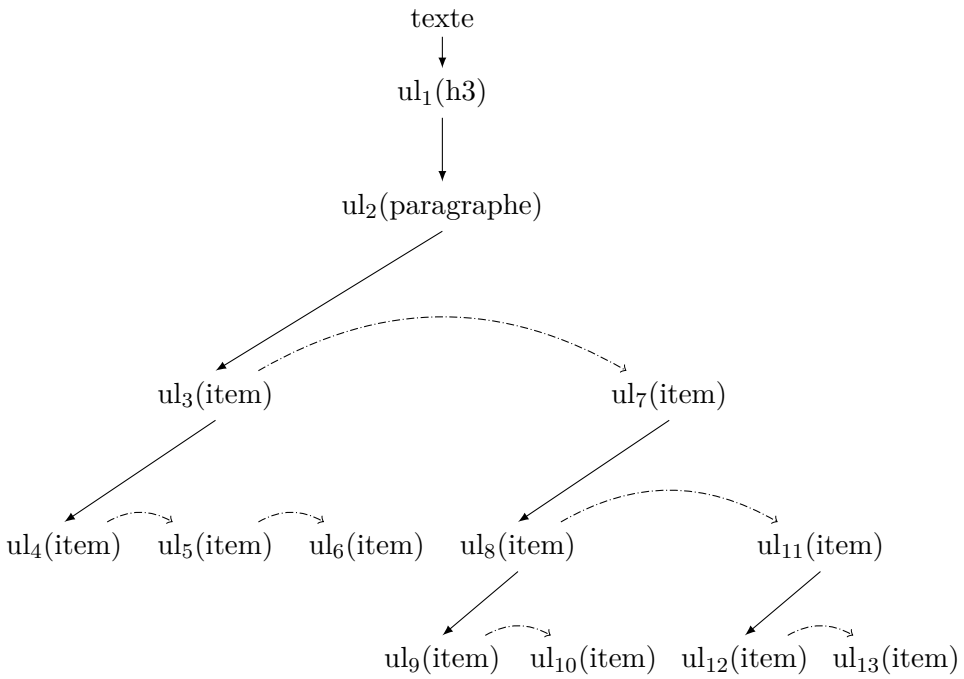

FIGURE 5.5 : Annotation en arbre de dépendances pour l'extrait de ling_muller

5.2 Segmentation en blocs textuels

Dans cette section, nous donnons la description de la méthode pour la segmentation des documents PDF en blocs textuels.

5.2.1 Description

La segmentation en blocs textuels repose sur l'utilisation d'un outil d'analyse géométrique. Dans ce travail, nous utilisons l'outil LA-PDFText, déjà utilisé pour l'annotation visuelle (Section 5.1.1). D'autres outils sont également envisageables si les informations que ceux-ci fournissent en sortie sont de natures dispositionnelle et typographique, et sont utilisables dans les deux tâches suivantes (Sections 5.3 et 5.4).

À ce stade de notre travail, la mise en œuvre de la tâche de segmentation en blocs textuels ne constitue pas une contribution directe. Pour l'évaluation de LA-PDFText, nous renvoyons à la publication correspondante (Ramakrishnan *et al.*, 2012).

Notons, toutefois, que dans le cadre d'un stage étudiant (terminé en octobre 2015), une large partie du code de LA-PDFText a été relue, factorisée et améliorée. Une interface a également été ajoutée. Le code source est librement accessible[13]. Ce travail a permis d'améliorer la tâche de segmentation en blocs textuels. Les résultats préliminaires obtenus sont discutés dans les perspectives de ce travail de thèse.

[13] http://github.com/fauconnier/lapdftext

5.3 Étiquetage automatique des blocs textuels en unités logiques

Dans cette section, nous donnons la description de la méthode. Ensuite, nous reportons l'évaluation obtenue sur la structure logique de surface annotée.

5.3.1 Description

Nous décrivons ici une méthode pour l'étiquetage des blocs textuels issus de la segmentation faite par un outil d'analyse géométrique. Les étiquettes logiques sont celles décrites en section 5.1.2. Nous considérons ce problème comme un problème d'apprentissage séquentiel (Daumé III, 2006) : l'hypothèse est faite que, pour un bloc textuel donné, (i) les indices visuels locaux sont informatifs, mais également (ii) les étiquettes précédemment apparues. Ces informations sont prises en compte par des traits (pour *features*). Deux points sont décrits dans cette section :

- la formalisation du problème en un problème d'apprentissage séquentiel ;
- les traits utilisés dans l'apprentissage séquentiel.

Apprentissage séquentiel Nous formalisons le problème d'étiquetage comme un problème d'apprentissage séquentiel. L'apprentissage séquentiel est un type d'apprentissage supervisé qui consiste à généraliser statistiquement des séquences d'observations[14].

Dans ce travail, nous avons choisi de représenter les blocs textuels des documents au travers de séquences de forme :

$$\mathbf{x} = (x_1, x_2, \ldots, x_m)$$

où chaque x_i représente un bloc textuel. Nous représentons les étiquettes logiques au travers de séquences de forme :

$$\mathbf{y} = (y_1, y_2, \ldots, y_m)$$

où y_i est l'étiquette logique associée au bloc textuel x_i. L'objectif est de trouver la séquence d'étiquettes logiques adéquates pour une séquence de blocs textuels donnée. Cette fonction cible a la forme :

$$f(\mathbf{x}) = \mathbf{y}$$

Comme il n'est pas possible de définir f analytiquement, nous l'approximons par une fonction (dite hypothèse) apprise à partir des exemples dans les corpus LING et GEOP. Dans ce travail, nous utilisons des Champs Conditionnels Aléatoires (*Conditional Random Fields* pour CRF) pour cet apprentissage. Les CRF, proposés par Lafferty *et al.* (2001), sont des modèles discriminants et probabilistes. Ceci leur confère deux avantages :

[14] Notons qu'il existe également des algorithmes d'apprentissage adaptés à d'autres types de structures. Citons, par exemple, XCRF (Jousse *et al.*, 2006) pour les arbres ou MIRA (McDonald *et al.*, 2005a) pour les graphes.

1. L'avantage des modèles discriminants sur les modèles génératifs (p. ex. Modèle de Markov Caché (Rabiner, 1989)) est qu'ils relâchent l'hypothèse d'indépendance entre les observations de la séquence[15]. Ceci permet de faire dépendre les transitions entre les étiquettes logiques par des observations précédemment vues. Il devient possible de capturer des régularités telles que, par exemple, un titre est généralement suivi d'un paragraphe.

2. L'avantage des modèles probabilistes sur les modèles non probabilistes, tels que les algorithmes cherchant un hyperplan séparateur (p. ex. perceptron structuré (Collins, 2002)), apparaît dans les situations où il n'existe pas de traits prédicateurs suffisamment forts pour déterminer avec exactitude l'étiquette d'un exemple donné, mais plutôt un faisceau d'indices. Dans ce cas, une estimation probabiliste est généralement préférable[16].

Ces avantages font des CRF des modèles adaptés pour notre tâche d'étiquetage.

Nous donnons ci-dessous une brève description formelle des CRF. Une explication plus complète est donnée dans l'annexe consacrée aux algorithmes d'apprentissage supervisé (Annexe B). Les CRF présentent une forme comparable à celle de la régression logistique multinomiale, c'est-à-dire une forme exponentielle, mais en la généralisant à une séquence. Ainsi, la normalisation est faite sur l'ensemble des séquences d'étiquettes possibles. Pour une séquence $\mathbf{y}$ fixe, les CRF ont la forme :

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(\theta^{\mathrm{T}} F(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(\theta^{\mathrm{T}} F(\mathbf{x}, \mathbf{y}'))} \qquad \text{(eq.5.1)}$$

Le vecteur θ est le vecteur de poids associés aux traits. Pour simplifier la notation, nous représentons artificiellement les traits par des fonctions[17]. La fonction $F(\mathbf{x}, \mathbf{y})$ retourne un vecteur global de traits. Chaque dimension $F^{(j)}$ est la somme sur la séquence $y_1, \ldots, y_m$ du trait local $f^{(j)}$ correspondant :

$$F^{(j)}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} f^{(j)}(\mathbf{x}, \mathbf{y}, i) \qquad \text{(eq.5.2)}$$

La fonction de trait local f représente les traits utilisés.

[15] Les modèles génératifs modélisent la distribution jointe des observations et des étiquettes $p(\mathbf{x}, \mathbf{y})$, tandis que les modèles discriminants modélisent uniquement la distribution conditionnelle des étiquettes $p(\mathbf{y}|\mathbf{x})$.

[16] *A contrario*, les algorithmes d'apprentissage à hyperplan séparateur donnent une décision fiable pour une observation uniquement si cette observation est suffisamment éloignée de cet hyperplan. Or, pour les problèmes complexes, cela n'est pas systématique.

[17] Cette position est notamment celle de Sutton et McCallum (2006) : « Rather than using one weight vector per class, (...), we can use a different notation in which a single set of weights is shared across all the classes. The trick is to define a set of feature functions that are nonzero only for a single class. To do this, the feature functions can be defined as $f_{y',j}(y, \mathbf{x}) = 1_{\{y'=y\}} x_j$ for the feature weights and $f_{y'}(y, \mathbf{x}) = 1_{\{y'=y\}}$ for the bias weights. »

Traits utilisés Chaque trait local f correspond à une information capturée dans les blocs textuels. Il existe deux catégories de traits locaux (Sha et Pereira, 2003) :

- **Traits d'états** Les traits d'états visent à donner des informations sur chaque état de la séquence pris séparément et prennent, à l'état i, la forme $f(\mathbf{x}, y_i, i)$.

  Le tableau 5.5 synthétise les traits d'états utilisés dans l'étiquetage des blocs textuels. Un travail de généralisation des traits *t_marges*, *t_fontes* et *t_position_(h/v)* a été effectué afin d'utiliser des valeurs propres à chaque document, et non au corpus. Ceci permet d'extraire des informations relatives (p. ex. une fonte apparaît majoritairement dans le document, il y a une indentation à gauche, etc.) à la place des valeurs absolues (p. ex. une fonte de taille 10, un retrait de 40 pixels, etc.). Cette manière de procéder permet de se détacher en partie des variations de mise en forme et d'éviter d'induire des biais dans l'apprentissage.

  Ce travail de généralisation des traits a nécessité une phase de pré-traitement pour chaque document. Cette phase calcule le mode des variables discrètes (p. ex. marges, tailles des fontes, etc.) et nominales telles que les types de police (p. ex. Times New Roman, Arial, etc.).

  Les traits de *t_typographie* signalent les blocs textuels commençant par un symbole éventuellement suivi d'un mot (p. ex. 1, *, ., -aab, 2aba, etc.). Les traits *t_ratios* résument plusieurs caractéristiques locales au sein d'une même valeur (p. ex. surface divisée par nombre de mots fois taille de la fonte). Les distributions de valeurs de ces ratios sont discrétisées au travers d'un découpage en quartiles, par document. À chaque quartile est associée une fonction booléenne qui renvoie vrai si le ratio du bloc textuel courant appartient à ce quartile.

Traits d'états	Informations capturées
t_marges	Indentation à droite ou à gauche, centrage des blocs, absence d'indentation, etc.
t_fontes	Présence d'emphases (gras ou italique), taille de la fonte, etc.
t_position_v	Position verticale dans la page (haut, bas) et horizontale (droite, gauche).
t_position_h	Position dans la séquence du document (début, fin)
t_typographie	Présence d'un symbole typographique en début de bloc textuel.
t_ratios	Ratios de la longueur sur la largeur, de la surface sur la taille de la fonte, etc.

TABLE 5.5 : Traits d'états pour l'étiquetage logique des blocs textuels

- **Traits de transitions** Les traits de transitions capturent les dépendances entre les blocs textuels de la séquence. Pour des raisons de complexité algorithmique, nous utilisons des CRF de premier ordre, c'est-à-dire où seul l'état précédent est pris en compte pour déterminer la transition qui suit. Ces traits de transitions prennent à l'état i la forme $f(\mathbf{x}, y_i, y_{i-1}, i)$.

Le tableau 5.6 synthétise les types de traits utilisés et les informations capturées. Les *t_bigrammes* capturent des paires d'étiquettes et les régularités de transitions associées (p. ex. les références bibliographiques se suivent généralement). Les *t_contrastes* comparent les fontes entre deux blocs textuels contigus.

Traits de transitions	Informations capturées
t_bigrammes	Étiquette attribuée au bloc textuel qui précède dans la séquence du document.
t_contraste	Rupture de fonte avec le bloc textuel qui précède (taille de fonte, type de police).

TABLE 5.6 : Traits de transitions pour l'étiquetage logique des blocs textuels

5.3.2 Évaluation

Dans cette section, nous donnons l'évaluation obtenue sur l'annotation de la structure logique de surface des corpus LING et GEOP (section 5.1.2). Nous avons procédé à une validation croisée (k =10) et présentons les résultats en termes d'exactitude.

Pour la comparaison, nous proposons une baseline simple consistant à étiqueter tous les blocs textuels en paragraphes, qui forment la classe majoritaire dans LING et GEOP. Pour chaque corpus, nous avons effectué l'évaluation selon deux configurations : (i) avec les *traits d'états* seuls et, ensuite, (ii) avec les traits d'états adjoints aux *traits de transitions*. Les résultats sont reportés dans le tableau 5.7.

Configurations	LING	GEOP	LING_GEOP
Traits d'états	79,97	81,09	74,58
+ Traits de transitions	**87,24**	**82,88**	**80,23**
baseline	32,33	44,51	37,33

TABLE 5.7 : Exactitude pour l'étiquetage en unités logiques élémentaires

Dans la première configuration, les résultats pour LING montrent une difficulté à étiqueter les blocs textuels, avec un léger recul par rapport à GEOP (Δ1,12%). Ce taux plus bas s'explique notamment par les nombreux exemples linguistiques au sein de LING (section 5.1.2). Ces blocs textuels, considérés comme items dans la structure logique de surface présentent des caractéristiques visuelles différentes des items « classiques ».

Dans la deuxième configuration, avec les traits de transitions, la prise en compte de la séquence permet de pallier en partie les variations locales des unités. Cela se traduit par des augmentations significatives (test t pour échantillons appariés avec un risque α à 0,05) par rapport à la première configuration dans LING ($t = 13{,}79$ et $p < 0{,}01$) et GEOP ($t = 3{,}10$ et $p < 0{,}02$). Toutefois, pour GEOP, cette amélioration (Δ1,79%) n'est pas aussi élevée que pour LING (Δ7,27%).

Dans les figures 5.6 et 5.7, nous donnons les F_1-scores obtenus par étiquettes. Le caractère académique et normé visuellement du corpus LING apparaît nettement dans les résultats. Les blocs textuels relatifs aux articles scientifiques (titre du document, byline, référence bibliographique, h1, h2, h3) sont relativement bien marqués et, par conséquent, plus facilement reconnus. L'étiquetage des items est relativement correct (F_1-score de 67,79). Toutefois, malgré la prise en compte des transitions, ceux-ci restent confondus avec les paragraphes, les citations et, enfin, les références bibliographiques. La convention du CMLF implique qu'il y a très peu d'en-têtes, de pieds de page ou de notes de bas de page (remplacées par des notes de fin de document). Les scores obtenus pour ces étiquettes sont nuls. La situation inverse apparaît pour GEOP.

L'hétérogénéité visuelle de GEOP complexifie l'étiquetage des blocs textuels. Cela apparaît notamment pour la titraille qui, bien que présentant une distribution équivalente à celle de LING, obtient des scores plus bas. Les nombreux blocs textuels de type *autres* (p. ex. légende de figure, contenu textuel de tableau, etc.) rompent régulièrement la séquence d'étiquettes dans GEOP, compliquant l'étiquetage. Ceci explique en partie l'efficacité moindre des traits de transitions pour ce corpus

Dans les deux corpus, les paragraphes, largement majoritaires, sont correctement étiquetés : F_1-scores de 90,95 pour LING et de 92,17 pour GEOP.

Afin de diminuer les variations de distribution dans les corpus d'apprentissage, une évaluation a été faite sur les deux corpus pris conjointement. Sur ce corpus nommé LING_GEOP, les traits de transitions montrent également un gain (Δ5,65%) par rapport aux traits d'états (Table 5.7). Une analyse par étiquettes, pour ce corpus, montre une amélioration globale (Figure 5.8). La figure 5.9 présente les courbes d'apprentissage avec les traits de transitions sur les trois corpus. Pour obtenir ces courbes, une validation croisée (k=10) à été exécutée pour chaque nombre n de documents choisis aléatoirement parmi les 9 ensembles restants. Ces résultats indiquent qu'un agrandissement du corpus améliore les scores.

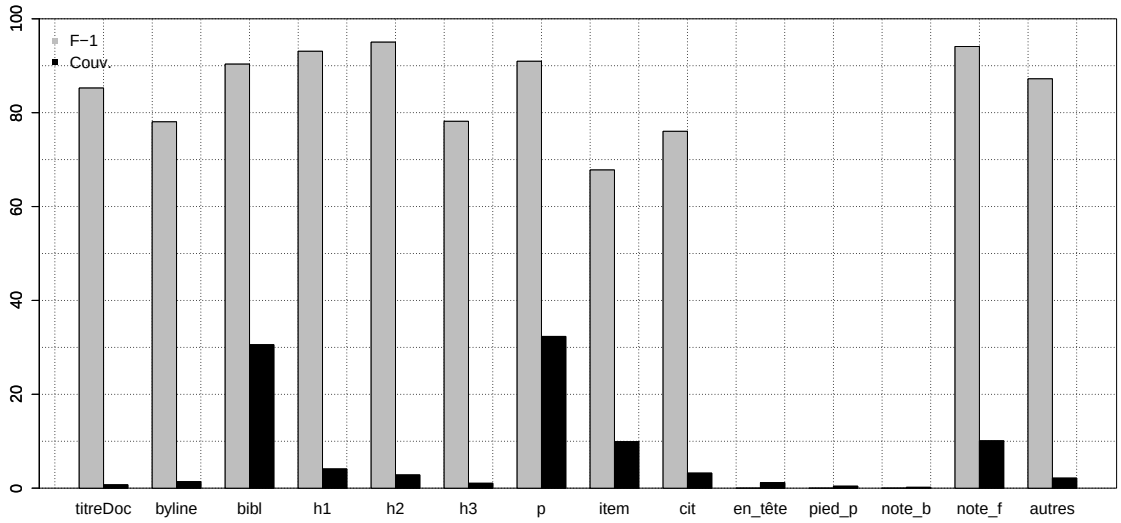

FIGURE 5.6 : F_1-scores pour les étiquettes avec leur couverture pour LING

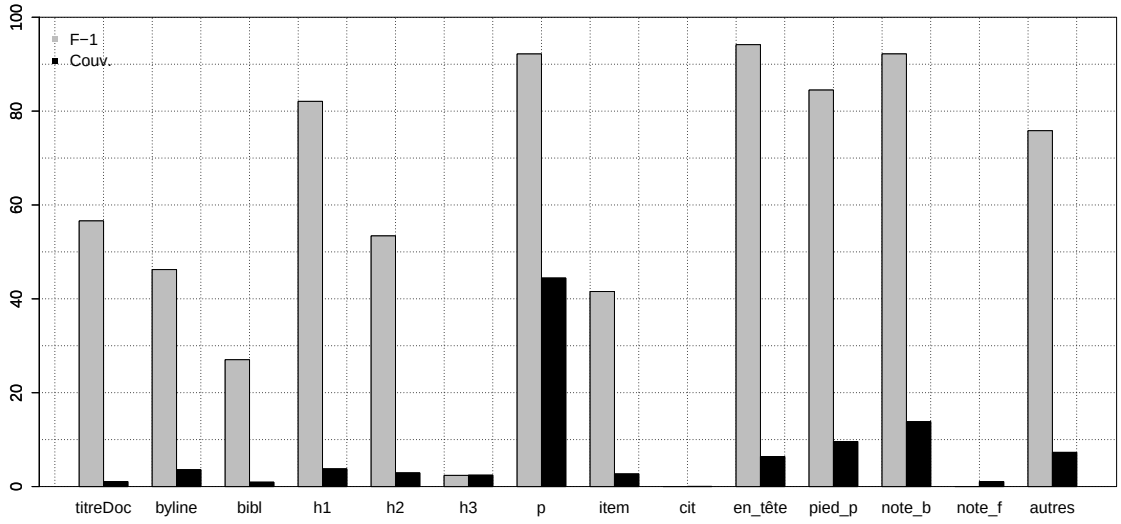

FIGURE 5.7 : F_1-scores pour les étiquettes avec leur couverture pour GEOP

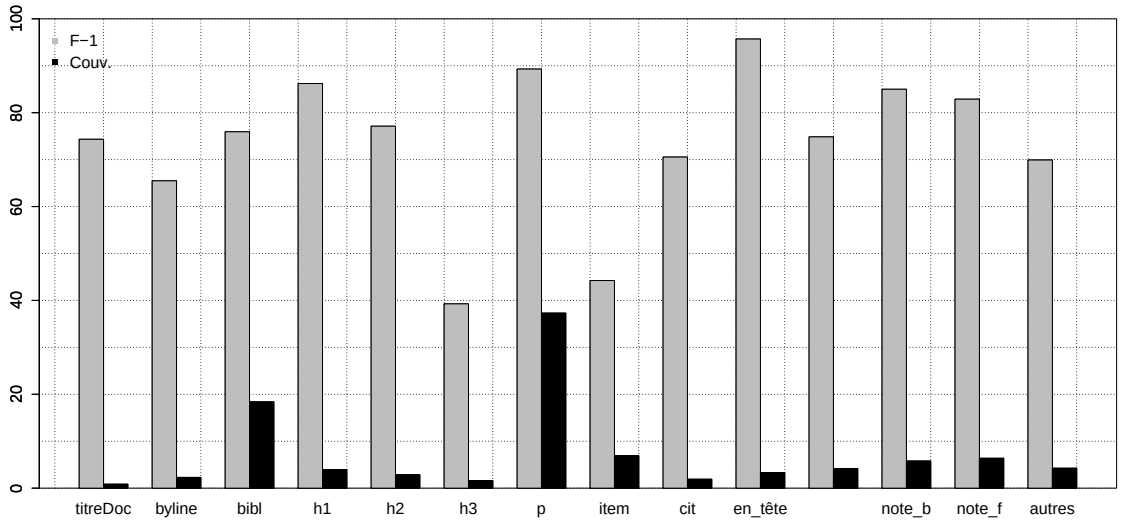

FIGURE 5.8 : F_1-scores pour les étiquettes avec leur couverture pour LING_GEOP

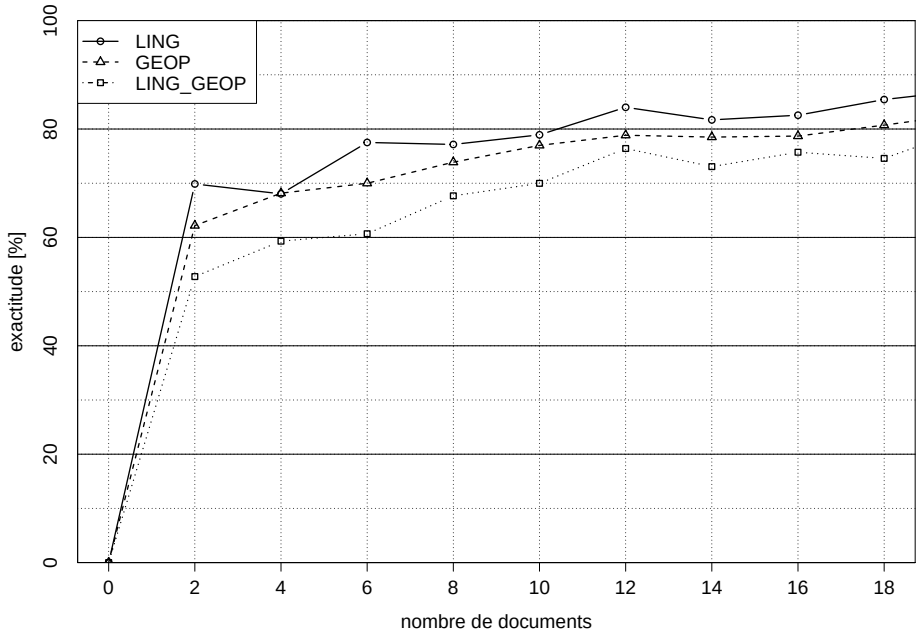

FIGURE 5.9 : Courbes d'apprentissage pour LING, GEOP et LING_GEOP

5.4 Représentation du document sous la forme d'un arbre de dépendances

Dans cette section, nous décrivons la méthode de construction de l'arbre de dépendances. Ensuite, nous reportons l'évaluation obtenue sur la structure logique profonde annotée.

5.4.1 Description

Pour cette tâche, deux objectifs sont poursuivis : (i) attribuer une relation de dépendance entre les unités logiques élémentaires et (ii) construire l'arbre correspondant à l'ordonnancement de ces unités au sein de chaque document. Les propriétés de notre représentation en dépendances (Chapitre 4) offrent la possibilité de réaliser ces deux objectifs simultanément avec des techniques comparables à celles utilisées en analyse syntaxique et en parsing rhétorique.

Dans notre travail, nous utilisons un parseur de type *shift-reduce*. À un niveau phrastique, ce type d'algorithme a déjà été utilisé pour la construction d'arbres rhétoriques (Marcu, 1999; Choi, 2002). En particulier, nous reprenons la version de Hernandez et Grau (2005) qui prend en compte la relation de coordination et nous l'avons adaptée au niveau de la structure du document[18]. Nous distinguons plusieurs composants au système de parsing de dépendances que nous proposons : (i) un algorithme de parsing et (ii) une méthode qui transpose les états du parseur en des actions de parsing[19]. Deux méthodes différentes sont proposées pour transposer les états du parseur en actions : la première repose sur une grammaire de dépendances et la seconde sur de l'apprentissage supervisé. La figure 5.10 schématise le système de parsing dans son ensemble.

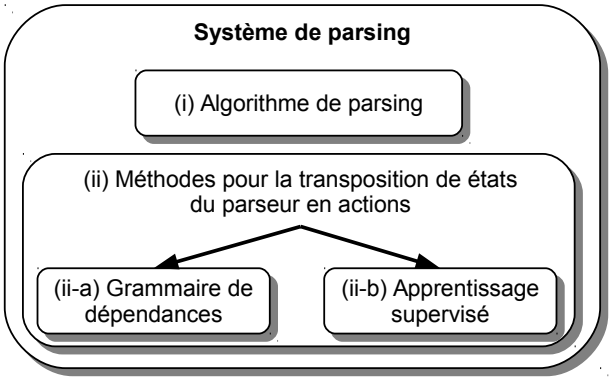

FIGURE 5.10 : Schéma du système de parsing pour la construction de l'arbre de dépendances

[18] La granularité des phénomènes étudiés par notre modèle a déjà été discutée dans le chapitre 4.
[19] Cette distinction à deux composants de notre système de parsing s'inspire de la distinction faite par Nivre (2008) pour le parsing syntaxique.

Dans cette section, nous allons successivement décrire chacun de ces composants :
(i) l'algorithme de parsing ;
(ii) les méthodes de transposition des états du parseur en actions :
(ii-a) la grammaire de dépendances ;
(ii-b) l'algorithme d'apprentissage avec ses traits.

(i) Algorithme de parsing L'algorithme de parsing utilisé est un analyseur LR(1)[20], s'inscrivant dans la famille plus générale des parseurs *shift-reduce*. Cet algorithme à pile parcourt la séquence des unités logiques élémentaires, de gauche à droite, en cherchant le point d'attachement optimal à gauche pour chaque unité logique, jusqu'à ce que la séquence soit réduite[21]. À chaque étape de la construction de l'arbre, l'unité courante ainsi que l'unité entrante sont prises en compte pour déterminer l'action à produire. Trois actions sont permises :

1. **reduce_subordination** L'unité courante est liée par une dépendance de type subordination tournée vers l'unité entrante. Ceci équivaut à une descente dans la hiérarchie du document.

2. **reduce_coordination** L'unité courante est liée par une dépendance de type coordination tournée vers l'unité entrante. Ceci équivaut à rester au même niveau dans la hiérarchie du document.

3. **shift** L'unité courante est mise de côté et remplacée par l'unité entrante. Ceci équivaut à une remontée dans la hiérarchie du document.

Dans l'algorithme 1, nous décrivons le procédé de parsing en pseudo-code. La fonction empiler(u, p) consiste à mettre l'unité logique u sur le dessus de la pile p. La fonction dépiler(p) consiste à enlever la dernière unité logique entrée dans la pile p. La fonction défiler(f) consiste à enlever l'unité logique de la position terminale de la file f et la retourner. La fonction choisir_action(u, v) prend en entrée les unités logiques u et v et appelle une des méthodes de transposition. Kübler *et al.* (2009) divisent ces méthodes en deux familles : (a) les approches basées sur des grammaires formelles (*grammar-based*) et (b) les approches conduites sur des données (*data-driven*), généralement statistiques. Nous comparons les deux méthodes dans ce travail.

Avant d'aborder ces méthodes, nous donnons un exemple simple de parsing de document. Nous posons ici que les titres subordonnent toujours les paragraphes dans la séquence du document. Additionnellement, les contraintes syntaxiques décrites dans le chapitre 4 s'appliquent. L'exemple porte sur la réduction de la séquence suivante :

$$d = (\text{ul}_1(\text{h1}), \text{ul}_2(\text{paragraphe}), \text{ul}_3(\text{paragraphe}), \text{ul}_4(\text{h1}), \text{ul}_5(\text{paragraphe})) \qquad \text{(eq.5.3)}$$

Les différentes étapes de réduction de cette séquence sont données dans la table 5.8. L'arbre de dépendances résultant est donné en figure 5.11.

[20] Analyseur LR(1) désigne un analyseur *Left to Right* avec la prise en compte d'une unité entrante en plus de l'unité courante. Se référer au travail de Knuth (1965).
[21] Autrement dit, le parsing s'arrête lorsqu'il n'est plus possible d'avoir de dérivations à droite.

Algorithme 1 Analyseur LR(1) adapté à la subordination et la coordination. σ est une pile d'unités logiques et β est une file d'unités logiques. Un sous-script est utilisé pour indiquer la position d'une unité logique dans la pile ou dans la file : σ_0 désigne l'unité du dessus de la pile et β_0 désigne la première unité de la file. Avant que débute le parsing, σ est vide et β contient la séquence des unités logiques du document. Les méthodes sur ces structures de données portent les noms des primitives associées. *texte* est une unité logique factice endossant le rôle de la racine de l'arbre de dépendances. D est l'ensemble des arcs dirigés où chaque arc est représenté par un triplet (u, t, v) où u et v sont des unités logiques et t appartient à l'ensemble $T = \{sub, coord\}$.

empiler(*texte*, σ)
Tant Que σ et β non vides **Faire**
    decision ← choisir_action(σ_0, β_0)
    Si *decision* == subordination : `//reduce_subordination`
        $D \leftarrow D \cup (\sigma_0, sub, \beta_0)$
        empiler(défiler(β), σ)
    Sinon Si *decision* == coordination : `//reduce_coordination`
        $D \leftarrow D \cup (\sigma_0, coord, \beta_0)$
        dépiler(σ)
        empiler(défiler(β), σ)
    Sinon `//shift`
        dépiler(σ)
    Fin Si
Fin Tant Que

	action	pile σ	file β	triplet ajouté à D
0		*texte*	h1, p, p, h1, p	
1	reduce_subordination	*texte*, h1	p, p, h1, p	(*texte*, *sub*, h1)
2	reduce_subordination	*texte*, h1, p	p, h1, p	(h1, *sub*, p)
3	reduce_coordination	*texte*, h1, p	h1, p	(p, *coord*, p)
4	shift	*texte*, h1	h1, p	
5	reduce_coordination	*texte*, h1	p	(h1, *coord*, h1)
6	reduce_subordination	*texte*, h1, p	∅	(h1, *sub*, p)
7	shift	*texte*, h1	∅	
8	shift	*texte*	∅	
9	shift	∅	∅	

TABLE 5.8 : Étapes de réduction de la séquence (eq.5.3). Pour la clarté de l'exemple, seules les étiquettes logiques sont reportées et l'étiquette 'paragraphe' est représentée par la lettre 'p'.

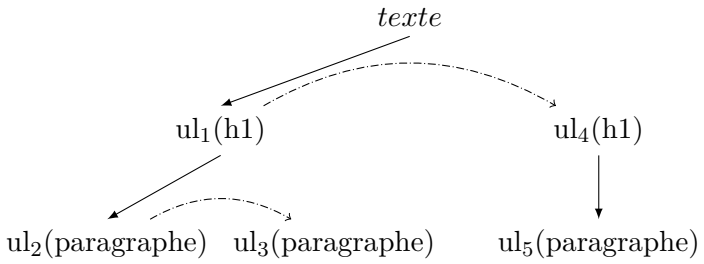

FIGURE 5.11 : Arbre de dépendances obtenu pour la réduction en table 5.8

(ii) Méthodes de transposition des états du parseur en actions

Nous proposons deux méthodes pour transposer les états du parseur en actions de parsing. La première méthode repose sur la définition d'une grammaire de dépendances et la seconde repose sur de l'apprentissage supervisé.

(ii-a) Grammaire de dépendances Nous utilisons une série de règles simples pour déterminer les relations entre les unités logiques. Ces règles se basent uniquement sur les informations fournies par les étiquettes logiques. L'ensemble de ces règles permet la construction d'une grammaire de dépendances[22].

Pour formaliser notre grammaire de dépendances, nous avons pris comme point de départ la catégorisation faite par Hellwig (2006). Cette catégorisation distingue quatre types de formalismes (Table 5.9). Les formalismes G1 et G3 concernent les grammaires de constituants, dont les représentants prototypiques sont la Grammaire Générative et Transformationnelle de Chomsky (1957), connue pour ses règles de réécriture de forme $S \rightarrow SN\ SV$, et les Grammaires Catégorielles (Bar-Hillel *et al.*, 1960). Les formalismes G2 et G4 concernent les grammaires de dépendances. Le type G2 se rapporte aux grammaires de dépendances exprimées par des règles. Les travaux de Hays (1964) et Gaifman (1965) rentrent dans cette catégorie. Enfin, le type G4 concerne les grammaires de dépendances uniquement basées sur la valence des mots.

Type de grammaire	à base de règles	à base de lexique
grammaire de constituants	G1	G3
grammaire de dépendances	G2	G4

TABLE 5.9 : Quatre types de formalismes pour les grammaires selon Hellwig (2006)

Nous reprenons ici le formalisme de Gaifman (1965), de type G2. En substance, celui-ci définit plusieurs ensembles de règles pour la formalisation d'un système d'analyse en dépendances[23]. Nous nous basons sur son formalisme d'expression des dérivations et

[22] La grammaire décrite ici est celle qui a été utilisée pour l'annotation semi-manuelle du corpus PDF (Section 5.1.3).

[23] « By a dependency system we mean a system, containing a finite number of rules, by which dependency

l'adaptons pour prendre en compte explicitement la subordination et la coordination[24]. Dans ce contexte, nous définissons que les règles de forme :

$$\mathrm{ul}_i(a)\ \mathrm{S}\Big(*\ \mathrm{ul}_j(b)\Big)$$

expriment que les unités d'étiquette logique a subordonnent les unités d'étiquette logique b. Le symbole * exprime la position que prend ul_i dans la dérivation. Dans notre cas, le symbole * sera toujours à gauche car seules les transitions à droite sont permises (section 4.3.1). Les règles de forme :

$$ul_i(a)\ \mathrm{C}\Big(*\ \mathrm{ul}_j(b)\Big)$$

expriment que les unités d'étiquette logique a coordonnent les unités d'étiquette logique b. Le soulignage signale les unités logiques qui peuvent être la racine de la séquence. Dans notre cas, seule l'unité logique factice *texte* (ul_0) peut être racine. Nous donnons ci-dessous un extrait des règles construisant notre grammaire hors contexte :

$$\underline{texte}\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{h1})\Big)$$

$$\underline{texte}\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{h2})\Big)$$

$$\underline{texte}\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{h3})\Big)$$

...

$$\mathrm{ul}_i(\mathrm{h1})\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{h2})\Big)$$

$$\mathrm{ul}_i(\mathrm{h1})\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{h3})\Big)$$

$$\mathrm{ul}_i(\mathrm{h1})\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{paragraphe})\Big)$$

$$\mathrm{ul}_i(\mathrm{h1})\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{item})\Big)$$

$$\mathrm{ul}_i(\mathrm{h1})\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\mathrm{citation})\Big)$$

$$\mathrm{ul}_i(\mathrm{h1})\ \mathrm{S}\Big(*\ \mathrm{ul}_j(\text{référence bibliographique})\Big)$$

...

analysis for a certain language is done (...) In the explication given here, this consists of the following three sets of rules : 1. Rules which give for each category those categories which may derive directly from it with their relative positions. (...) 2. Rules giving for every category the list of all words belonging to it. (...) 3. A rule giving the list of all categories the occurrence of which may govern a sentence. » (Gaifman, 1965, p. 305 - 306)

[24] La plupart des formalismes adaptés aux langues naturelles représentent implicitement la subordination et la coordination par la forme des arbres.

$\text{ul}_i(\text{paragraphe})\ \text{S}\big(*\ \text{ul}_j(\text{item})\big)$

$\text{ul}_i(\text{paragraphe})\ \text{S}\big(*\ \text{ul}_j(\text{citation})\big)$

...

$\text{ul}_i(\text{paragraphe})\ \text{C}\big(*\ \text{ul}_j(\text{paragraphe})\big)$

$\text{ul}_i(\text{item})\ \text{C}\big(*\ \text{ul}_j(\text{item})\big)$

$\text{ul}_i(\text{citation})\ \text{C}\big(*\ \text{ul}_j(\text{citation})\big)$

...

Il est possible de généraliser les règles de subordination concernant la racine en posant la règle que *texte* subordonne toutes les étiquettes logiques sauf elle-même[25]. Également, les règles de coordination peuvent être généralisées par la règle $ul_i(a)\ \text{C}(*\ \text{ul}_j(b))$ si $a = b$.

La principale limite de cette méthode basée sur une grammaire de dépendances est qu'elle ne prend en compte que les étiquettes des unités logiques. Les indices visuels, tels que les retraits signalant une indentation graphique, ne sont pas pris en compte ici.

(ii-b) Apprentissage supervisé Nous utilisons un algorithme d'apprentissage supervisé pour déterminer la nature ou l'absence de dépendance entre deux unités logiques évaluées à un état donné du parseur. Le problème est considéré comme un problème de classification. Chaque état du parseur est représenté par un vecteur de traits $\mathbf{x}$. Les traits portent sur les informations visuelles, lexicales mais également sur les étiquettes logiques des deux unités logiques évaluées. L'algorithme doit associer cet état à une des classes parmi l'ensemble $Y = \{subordination, coordination, shift\}$. Ainsi, l'objectif est de trouver la classe adéquate pour un état donné. Cette fonction cible prend la forme :

$$f(\mathbf{x}) = y$$

Pour approximer f, nous avons choisi une régression logique multinomiale (Hastie *et al.*, 2009), aussi appelée classifieur d'entropie maximale (MaxEnt) (Berger *et al.*, 1996; Ratnaparkhi, 1996). Il s'agit d'une généralisation à $|Y|$ classes de la régression logistique binomiale (Cox, 1959).

L'intérêt de la régression logistique multinomiale est qu'elle permet une classification multi-classes. Il est possible d'utiliser des algorithmes de classification binomiale avec certaines stratégies telles que la *one-vs.-rest* ou la *one-vs.-one* (Bishop, 2006). Cependant, deux critiques peuvent être faites sur ces stratégies :

1. Ces stratégies estiment les paramètres de chaque classe de manière indépendante. Le système obtenu est alors davantage sensible aux données extrêmes, contrairement à une régression logistique multinomiale où les paramètres de chaque classe sont estimés de manière interdépendante.

[25] Il s'agit d'une contrainte syntaxique du modèle (Section 4.3.1).

2. Ces stratégies ne permettent pas d'avoir des probabilités en sortie, car se limitant généralement à construire $|Y| - 1$ modèles et à choisir la classe pour laquelle le score est maximisé.

Pour ces raisons, la régression logistique multinomiale nous paraît le modèle le plus adapté dans notre situation.

Nous donnons une brève description formelle de la régression logistique multinomiale. Une explication plus complète est donnée dans l'annexe consacrée à l'apprentissage supervisé (Annexe B). La régression logistique multinomiale est une généralisation de la régression logistique binomiale. Dans ce contexte, il est nécessaire de normaliser la distribution. Pour une classe y fixe, la régression logistique multinomiale a la forme :

$$p(y|\mathbf{x}) = \frac{\exp(\theta_y^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{|Y|}\exp(\theta_c^{\mathrm{T}}\mathbf{x})} \qquad \text{(eq.5.4)}$$

Sur le plan théorique, le problème dual de la régression logistique proposé par Berger *et al.* (1996) où il s'agit de choisir, sous des contraintes calculées à partir des traits, la distribution maximisant l'entropie, est semblable à celui des Champs Conditionnels Aléatoires (précédemment introduits) qui maximisent la somme des entropies de la séquence sous des contraintes calculées identiquement (Ganapathi *et al.*, 2008)[26].

Nous présentons ci-dessous les traits pour représenter les deux unités logiques évaluées à chaque état de parsing. Les traits utilisent (i) des informations visuelles (typographiques et dispositionnels), (ii) des informations lexicales, (iii) les étiquettes des unités logiques élémentaires et, enfin, (iv) des informations liées au parallélisme (visuel et lexical) entre unités logiques. Le tableau 5.10 présente synthétiquement ces traits.

Les marqueurs liés aux traits *t_visuels* sont obtenus de manière similaire à l'étiquetage (vus dans la section précédente). Les traits *t_lexique* utilisent une liste prédéfinie de marqueurs d'intégration linéaire. Les traits *t_étiquettes* et *t_parallélisme* reposent sur l'hypothèse que nous disposons des résultats produits en sortie de l'étiquetage logique.

Traits	Informations capturées
t_visuels	Présence d'indentation, de tirets, de puces, de « : », etc.
t_lexique	Présence de MIL (p. ex. soit, soit, etc.).
t_étiquettes	Paires d'étiquettes (p. ex. titre-paragraphe, item-item, paragraphe-item, etc.) et égalité d'étiquettes.
t_parallélisme	Paragraphe entre deux items visuellement identiques, deux items différents visuellement, etc.

TABLE 5.10 : Traits pour la construction de l'arbre de dépendances

[26] Cela s'explique car l'expression du problème dual des Champs Conditionnels Aléatoires est la même que celle du modèle Markovien de maximisation d'entropie (*Maximum-entropy Markov Model* pour MEMM) (McCallum *et al.*, 2000).

5.4.2 Évaluation

Dans cette section, nous présentons l'évaluation sur la structure logique profonde annotées dans les corpus LING et GEOP (Section 5.1.3). Une relation de dépendance est considérée comme juste lorsque que (i) l'attachement à l'unité logique *tête* est juste et (ii) lorsque le type de la dépendance est juste. Nous avons procédé à validation croisée (k =10) et présentons les résultats avec une métrique d'exactitude.

Les deux méthodes pour la transposition des états du parseur en actions de parsing sont évaluées : la grammaire de dépendances et l'algorithme d'apprentissage supervisé. Pour la comparaison, nous proposons une baseline simple consistant à classer aléatoirement les dépendances de subordination et de coordination liant deux unités logiques. Les résultats sont reportés dans le tableau 5.11.

Méthodes	LING	GEOP	LING_GEOP
Grammaire	**96,54**	98,30	97,08
Apprentissage	96,41	**98,45**	**97,23**
Baseline	40,21	41,03	39,79

TABLE 5.11 : Exactitude pour la construction de l'arbre de dépendances

Pour les deux méthodes, la différence des résultats obtenus sur LING et GEOP s'explique par la structuration complexe, en termes de dépendances, au sein de LING. Par exemple, certains documents dans LING montrent des niveaux d'imbrications profonds, notamment par l'utilisation d'exemples linguistiques imbriqués dans des énumérations de définitions.

Cette différence est plus marquée pour la dépendance de subordination (F_1-scores autour de 92 pour LING et F_1-scores autour 97 pour GEOP), tandis que les scores pour la coordination sont plus stables (avec une légère amélioration pour GEOP). Tous les F_1-scores ainsi que les scores de Précision et Rappel associés sont reportés dans le tableau (Table 5.12).

		LING			GEOP		
Méthodes	Dép.	Précision	Rappel	F_1-score	Précision	Rappel	F_1-score
Grammaire	Sub	93,77	90,62	92,17	98,42	95,40	**96,88**
	Coord	97,31	98,26	97,78	98,27	99,42	**98,84**
Apprentissage	Sub	92,25	91,74	91,99	98,43	95,91	**97,15**
	Coord	97,61	97,77	97,69	98,46	99,42	**98,94**

TABLE 5.12 : Scores de Rappel, Précision et F_1-scores obtenus pour les types de dépendances par méthodes et par corpus

De manière générale, les résultats ne montrent pas de différences significatives entre la méthode par grammaire de dépendances et celle par apprentissage supervisé. Deux raisons expliquent cela :

1. Les dépendances entre les unités logiques suivent majoritairement les règles définies dans la grammaire. Seuls certains cas où l'indentation marquée visuellement intervient (p. ex. imbrications multiples, etc.) permettent de distinguer la grammaire de dépendances et l'apprentissage supervisé.

2. Cette asymétrie dans la distribution des cas (respectent vs. ne respectent pas la grammaire) induit pour l'apprentissage supervisé un phénomène d'apprentissage de la grammaire elle-même et non des traits considérés comme discriminants (p. ex. deux items contigus visuellement différents, un paragraphe avec retrait, etc.).

Pour mesurer plus finement les différences entre les deux méthodes, nous proposons deux stratégies d'évaluation dont nous reportons les résultats dans la table 5.13. La première consiste à évaluer uniquement les cas où la dépendance entre deux unités logiques diffère de la grammaire, c'est-à-dire lorsque la grammaire ne peut fournir la réponse correcte. Les résultats de cette stratégie montrent un léger gain qui reste stable au travers des corpus. La seconde consiste à évaluer l'apprentissage supervisé hors de l'ensemble des cas erronés de la grammaire de dépendances. Les résultats obtenus montrent 20 erreurs pour LING, 2 pour GEOP et 12 pour LING_GEOP. Notons, encore une fois, que l'augmentation du corpus d'apprentissage semble améliorer les résultats.

Stratégies	LING	GEOP	LING_GEOP
Exactitude de l'apprentissage sur l'ensemble des dépendances mal classées par la grammaire	14,54% (16/110)	16,66% (4/24)	14,17% (19/134)
Exactitude de l'apprentissage sur l'ensemble des dépendances correctement classées par la grammaire	99,34% (3051/3071)	99,85% (1394/1396)	99,73% (4455/4467)

Table 5.13 : Comparaisons entre les méthodes d'apprentissage supervisé et de grammaire de dépendances

5.5 Discussion

Nous avons défini un modèle décrivant la structure logique des documents dans le chapitre 4, et nous avons décrit une méthode pour son implémentation dans ce chapitre 5.

L'objectif a été de permettre l'analyse logique de documents de manière automatique. Pour cela, il a été choisi de caractériser les blocs textuels des documents et un travail d'abstraction et de généralisation des indices visuels a été mené afin de pallier la variabilité de ceux-ci. Nous avons proposé une méthode ascendante en trois étapes : (i) segmentation en blocs textuels, (ii) étiquetage des blocs textuels et (iii) construction de l'arbre de dépendances. Pour le format PDF, la première étape repose sur un outil

d'analyse géométrique et les deux autres étapes ont été implémentées. Ces implémentations ont été évaluées sur des corpus de natures différentes : une mise en forme unifiée et une structure complexe dans le corpus LING, un formatage hétérogène et une structure linéaire dans le corpus GEOP. Les résultats obtenus pour les deux étapes sont corrects. Toutefois, des expériences consistant à utiliser en séquence les deux étapes ont montré que la construction de l'arbre de dépendances était très sensible au bruit. Ceci constitue pour l'instant un aspect limitatif de notre solution.

L'adoption d'une représentation de la structure logique sous la forme d'un arbre et l'usage de méthodes statistiques sont des éléments partagés avec certains travaux proposés dans la communauté d'Analyse du Document (Paaß et Konya, 2012; Palfray *et al.*, 2012). Néanmoins, ces travaux ne reposent pas sur un modèle complet d'expression de la structure logique, qui est généralement sous-définie. D'autres travaux ont utilisé conjointement mise en forme visuelle et contenu lexical (Klink *et al.*, 2000; Ratté *et al.*, 2007), mais sans proposer de solution traitant la récursivité des constructions textuelles (p. ex. structures imbriquées).

Nous avons vu qu'utiliser un arbre de dépendances à la place d'un arbre de constituants offrait des avantages en termes de flexibilité d'analyse et de manipulation de structures hiérarchiques. Dans ce cadre, les relations de dépendances ouvrent la voie à l'identification de phénomènes discursifs complexes marqués visuellement. La partie III de ce travail s'intéressera à un de ces phénomènes : les structures énumératives verticales.

Troisième partie

Extraction de relations dans les structures énumératives verticales

Chapitre 6

Typologie et annotation des structures énumératives

Sommaire

Dans ce chapitre, nous nous intéressons aux structures énumératives (SE) verticales. Leur statut de structure énumérative rend ces structures textuelles intéressantes, car elles sont souvent porteuses de relations hiérarchiques (Chapitre 3). Leur mise en forme verticale implique que celles-ci sont identifiables aisément dans l'arbre de dépendances représentant la structure hiérarchique du document (Chapitres 4 et 5).

Nous décrivons ici une typologie des structures énumératives. Celle-ci nous permet de caractériser et cibler les structures énumératives porteuses de relations sémantiques utiles à la construction de ressources. Quatre dimensions sont considérées : les dimensions visuelle, rhétorique, intentionnelle et sémantique. L'objectif a été de vérifier l'existence de liens entre ces dimensions. Notre typologie se veut complémentaire à celle de Luc (2000) (Section 3.2.1) et orthogonale à celle de Ho-Dac *et al.* (2010) (Section 3.2.2).

Nous avons utilisé notre typologie pour établir un schéma d'annotation. Celui a été utilisé dans une campagne d'annotation à trois annotateurs. Le corpus résultant nous a permis d'obtenir un retour quantitatif sur la typologie, mais également de mettre au jour des indices de nature typographique, dispositionnelle, lexicale et syntaxique, qui seront utiles dans la mise en œuvre du processus d'extraction de relations (Chapitre 7).

6.1 Typologie multi-dimensionnelle des structures énumératives

Dans cette section, nous décrivons une typologie multi-dimensionnelle pour la caractérisation des SE. Elle s'appuie sur les dimensions visuelle, rhétorique, intentionnelle et sémantique. Les différentes caractéristiques observées au sein de chacune de ces dimensions sont illustrées par des exemples extraits du corpus de Virbel (1999) et du corpus de Kamel et Rothenburger (2011).

6.1.1 Axe visuel

D'un point de vue visuel, la SE a la propriété de pouvoir être formulée de diverses façons. Elle peut être énoncée discursivement en dehors de toute mise en forme, au sein de la même phrase ou au travers de plusieurs phrases n'appartenant pas nécessairement au même paragraphe. Elle peut également être mise en évidence par l'usage d'indices typographiques et dispositionnels, — indices qui peuvent suppléer l'absence d'indices lexico-syntaxiques. Ces indices typographiques et dispositionnels sont très variables (Chapitres 4 et 5), et permettent d'organiser les composants de la SE qui ne sont pas forcément contigus.

Dans ce contexte, nous distinguons deux types visuels : la **SE horizontale** qui peut bénéficier ou non d'une mise en forme typographique et la **SE verticale** qui bénéficie d'une mise en forme typographique et dispositionnelle.

La **SE horizontale** s'inscrit dans la linéarité du texte et ne fait pas usage du dispositionnel. Elle est caractérisée soit par des marqueurs d'intégration linéaire comme « premièrement », « deuxièmement », « d'abord », « ensuite », etc. qui permettent d'introduire les items séparément (SE 6.a), soit par des marqueurs lexicaux comme « tels que », « comme », etc. qui permettent d'introduire l'énumération (SE 6.b). Notons que la SE horizontale peut également faire usage de marqueurs typographiques pour délimiter son énumération, comme les parenthèses dans la SE 6.c.

(6.a)	Deux phénomènes sont responsables de l'augmentation substantielle du rayon de l'étoile (qui peut atteindre un rayon 1 000 fois supérieur à celui du Soleil). Premièrement, la fusion en couche de l'hydrogène. Et deuxièmement, la contraction du cœur d'hélium, libérant une importante quantité d'énergie gravitationnelle.

(6.b)	Le dromadaire a été répertorié dans 35 pays, tels que l'Inde, la Turquie, le Kenya, le Pakistan, la corne de l'Afrique et bien d'autres encore.

(6.c) Les Grecs fabriquent généralement des meubles en bois (type érable, chêne, if, saule), mais aussi en pierre et en métal (bronze, fer, or, argent).

La **SE verticale** présente des discontinuités par rapport à la linéarité du texte. Des marqueurs typo-dispositionnels sont utilisés pour organiser, subdiviser et hiérarchiser les différents composants de la SE (SE 6.d). Dans ce cas, les items apparaissent généralement en retrait visuel par rapport à l'amorce et sont introduits par des puces, des tirets, etc.

(6.d) Une chaussure se compose principalement :
- du semelage, partie qui protège la plante des pieds, plus ou moins relevée à l'arrière par le talon
- de la tige, partie supérieure qui enveloppe le pied

Les SE verticales et horizontales peuvent être combinées et imbriquées au sein d'une même SE. C'est le cas lorsqu'un item amène lui-même une SE, avec ou sans mise en forme typo-dispositionnelle (SE 6.e).

(6.e) Le bénéfice imposable est la différence entre les recettes et les charges de l'entreprise durant l'exercice comptable.

- Sont pris en compte pour les produits (recettes) :
  - les produits d'exploitation autrement dit le chiffre d'affaires de l'entreprise ;
  - les produits accessoires, c'est-à-dire les recettes.
- Sont pris en compte pour les charges (...) retenues pour leur coût hors taxe :
  - les frais généraux : salaire, loyer commercial, frais de bureau, etc. ;
  - les charges financières (agios, intérêts d'emprunt)

6.1.2 Axe rhétorique

À ce niveau nous prenons en compte la nature des relations du discours qui relient les items de la SE. Dans le cadre de la *Rhetorical Structure Theory* (Mann et Thompson, 1988), ces relations peuvent être de type noyau-satellite ou multi-nucléaire. Une relation noyau-satellite relie une unité du discours plus saillante à une unité du discours supportant l'information d'arrière-plan, alors qu'une relation multi-nucléaire relie des unités du discours d'importance égale.

Nous distinguons quatre types rhétoriques : les **SE paradigmatiques**, les **SE syntagmatiques**, les **SE hybrides** et les **SE bivalentes**. Ceci rejoint la terminologie précédemment proposée par Luc (2001) dans sa typologie.

La **SE** est **paradigmatique** si son énumération est paradigmatique. Dans ce cas, l'énumération porte une relation rhétorique multi-nucléaire entre les items (Figure 6.1). Les SE 6.a, 6.b, 6.c, entre autres, sont des cas de SE paradigmatiques.

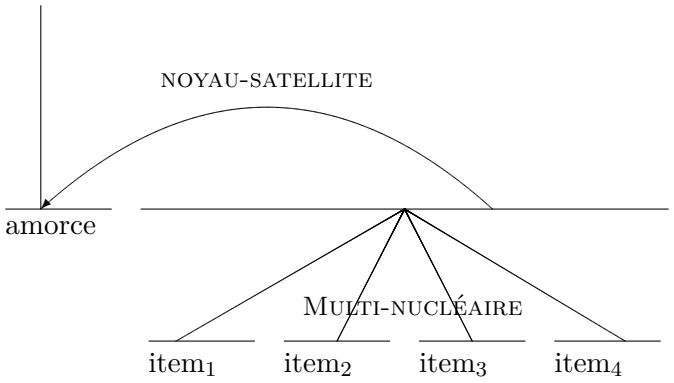

FIGURE 6.1 : Représentation rhétorique d'une SE paradigmatique

À l'opposé, la **SE syntagmatique** est composée d'une énumération portant des relations rhétoriques de type noyau-satellite entre ses items successifs (Figure 6.2). Les items ne sont donc pas interchangeables. La SE 6.f en est un exemple.

(6.f)

Est considéré comme « lecture savante », du point de vue fonctionnel, une pratique de lecture répondant aux critères suivants :

- c'est une lecture « qualifiée »,
- qui se développe sur le temps long de la recherche scientifique,
- dans un parcours forcément individualisé,
- où l'écriture se combine à la lecture, souvent dans une perspective de publications.

Lorsqu'une SE porte une relation rhétorique noyau-satellite entre au moins deux items et une relation rhétorique multi-nucléaire entre au moins deux items, elle est qualifiée d'**hybride**. Un exemple avait déjà été commenté dans le chapitre 3 (SE 3.e). Enfin, les caractères paradigmatique et syntagmatique peuvent coexister au sein de la même SE, et dans ce cas la SE est dite **bivalente** (SE 6.g).

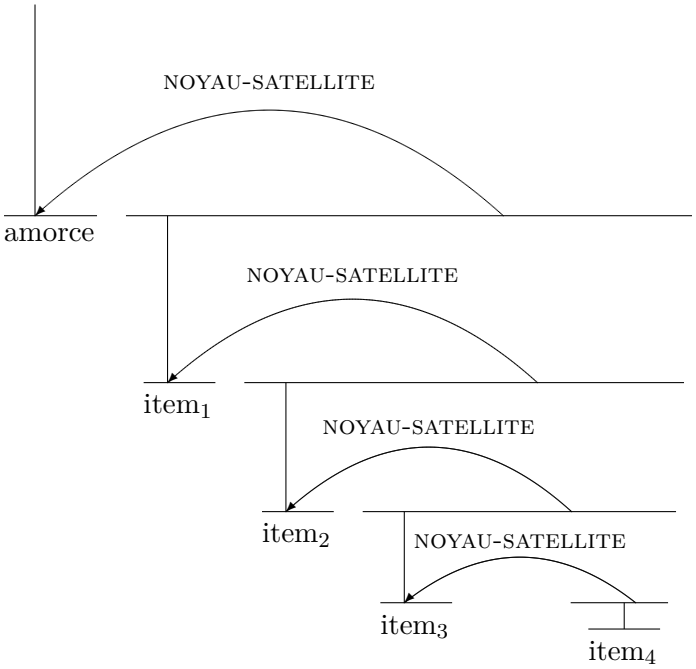

FIGURE 6.2 : Représentation rhétorique d'une SE syntagmatique

(6.g)

Chaque nucléotide est constitué de trois éléments liés entre eux :

- un groupe phosphate lié à :
- un sucre, le désoxyribose, lui-même lié à :
- une base azotée.

6.1.3 Axe intentionnel

À ce niveau nous prenons en compte l'intention de communication de l'auteur. Nous avons repris les types de textes d'Adam (2011)[1] pour les adapter aux SE, en différenciant les **SE descriptives**, les **SE narratives**, les **SE prescriptives**, les **SE procédurales**, les **SE explicatives**, et les **SE argumentatives**. Comme pour les axes précédents, l'objectif est de caractériser les types propices à la construction de ressources sémantiques.

La **SE descriptive** décrit une entité qui peut être un objet du monde animé ou pas, artificiel ou naturel (SE 6.a, 6.b, 6.c), alors que la **SE narrative** articule une succession d'actions ou d'événements, réels ou imaginaires (SE 6.h). Les notions de conseil, d'indication, d'injonction peuvent être intégrées à ces types de SE. Dans ce cas la SE est dite

[1] Ces types sont appelés séquences élémentaires prototypiques et permettent de classer les textes en types descriptif, narratif, explicatif, argumentatif, et dialogal (2011).

prescriptive (SE 6.i). De plus, lorsque ces conseils, indications, injonctions sont énoncés selon une volonté d'ordonnancer (comme dans les modes d'emploi, les notices explicatives, les guides d'utilisation, les manuels, les recettes de cuisine, etc.), pour atteindre un but donné, la SE est dite **procédurale** (SE 6.j).

(6.h)

Les Berbères ont mené une vive résistance parfois qualifiée de « farouche ».

- Algérie : De nombreux soulèvements ont été menés pour contrer la colonisation française, l'émir Abd el-Kader qui faisait remonter ses origines à la tribu berbère des Banou Ifren (Zénètes) a lutté après avoir déclaré la guerre aux Français, il fut capturé puis fait prisonnier. En juillet 1857, (...)
- Maroc : Le mouvement de résistance s'est illustré lors de la guerre du Rif menée par Abdelkrim al-Khattabi, qui est une guerre coloniale qui opposa les tribus berbères du rif aux armées françaises et espagnoles, de 1921 à 1926. (...)
- Libye : La lutte contre la colonisation italienne est d'abord menée par Omar Al Mokhtar surnommé « Cheikh des militants » qui est un chef musulman libyen d'origine berbère qui organisa la lutte armée contre la colonisation italienne au début du XXe siècle. D'autres leaders nationalistes (...)

(6.i)

Selon ce décret, la BnF a pour mission :

- de collecter, cataloguer, conserver et enrichir dans tous les champs de la connaissance, le patrimoine national dont elle a la garde, en particulier le patrimoine de langue française ou relatif à la civilisation française.
- d'assurer l'accès du plus grand nombre aux collections, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec la conservation de ces collections.

(6.j)

Préparation de la recette :

Lavez les asperges, épluchez-les de la pointe vers la base. Faites-les cuire dans une casserole d'eau bouillante avec les tablettes de bouillon pendant 25 à 30 minutes.

Égouttez-les et déposez-les précautionneusement sur du papier absorbant. Laissez-les refroidir.

Coupez-les en deux en réservant les pointes d'une longueur de 10 à 12 cm d'une part, les queues d'autre part.

La **SE explicative** répond en général à un questionnement de type « comment ? », « pourquoi ? », « dans quelles circonstances ? », etc. (SE 6.f). Si des arguments sont avancés dans le but de défendre une opinion, dans le but de convaincre, la SE est dite **argumentative** (SE 6.k).

(6.k) Du point de vue de la tradition textuelle juive, la division en chapitres est non seulement une innovation étrangère sans aucun fondement dans la messora, mais elle est également fort critiquable car :

- la division en chapitres reflète souvent l'exégèse chrétienne de la Bible ;
- quand bien même ce ne serait pas le cas, elle est artificielle, divisant le Texte en des endroits jugés inappropriés pour des raisons littéraires ou autres.

En ce qui concerne cet axe, une même SE pourra posséder plusieurs types intentionnels. La hiérarchie présentée en figure 6.1.3 décrit les combinaisons de types intentionnels rencontrées dans les corpus de Virbel (1999) et Kamel et Rothenburger (2011).

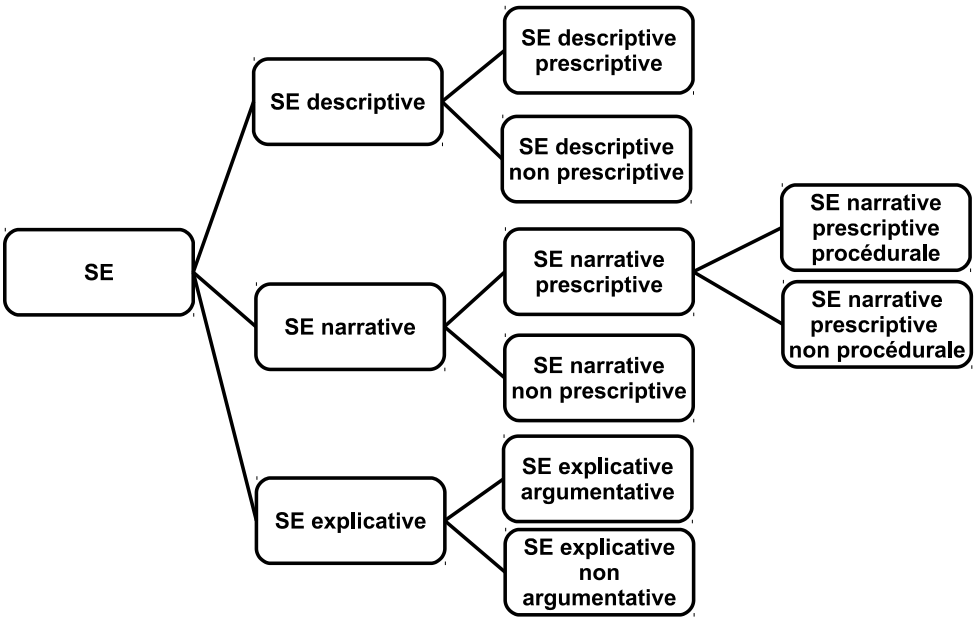

FIGURE 6.3 : Combinaisons rencontrées des types intentionnels au sein d'une même structure énumérative

Notons qu'il existe des SE pour lesquels aucun des types de l'axe intentionnel précités n'a pu être identifié. Pour celles-ci, nous avons établi le type **SE intentionnelle autre**.

6.1.4 Axe sémantique

D'un point de vue sémantique, les SE peuvent exprimer des connaissances de nature différente. Ces connaissances peuvent décrire de façon consensuelle ou conjoncturelle le monde réel ou imaginaire, la langue, les émotions, les sentiments, les opinions, etc. Nous rendons compte ici de cette dimension référentielle, conformément à notre objectif d'extraction de relations utiles à la construction de ressources sémantiques.

Nous avons divisé les SE en trois types sémantiques : **SE à visée ontologique** qui concerne des connaissances du monde (SE 6.a, 6.b, . . .), **SE metalexicale** qui concerne le fonctionnement du langage (SE 6.n, 6.o) et **SE sémantique autre** qui regroupe les SE qui ne sont ni à visée ontologique, ni metalexicale (SE 6.l).

(6.l)

S sait que p si et seulement si

1. p est vrai ;
2. S croit que p ; et
3. la croyance de S dans p est justifiée.

Les SE de type **à visée ontologique** peuvent être porteuses de relations hiérarchiques. Nous distinguons notamment l'hyperonymie (SE 6.a, 6.b, 6.c), la relation « instance de » (SE 6.m) et l'holonymie (SE 6.d, 6.g). Pour chacune de ces relations, nous définissons un sous-type sémantique de même nom. Les SE pour lesquelles aucune de ces relations ne s'appliquent, soit car la relation est difficile à définir, soit car la relation est une relation de domaine, sont associées au type **ontologique autre** (SE 6.i).

(6.m)

Manoirs célèbres

- Le manoir d'Ango à Varengeville-sur-mer, près de Dieppe.
- Le manoir de Brion au Mont-Saint-Michel
- Le manoir d'Eyrignac à Salignac-Eyvigues en Périgord

Les SE de type **metalexical** peuvent être porteuses des relations que Grabar *et al.* (2004) appellent des relations *lexicales*. Nous distinguons notamment la synonymie, l'antonymie, l'homonymie (SE 6.n) ou encore l'équivalence de traduction (SE 6.o). Le choix du terme *metalexical* est fait, car il nous semble que ces relations décrivent le langage par lequel elles sont elles-mêmes exprimées. Ceci fait écho aux réflexions faites sur le métalangage Harrissien (Section 2.1.3).

(6.n)
Une arête est un nom commun féminin qui peut désigner :
- l'arête, 'barbe de l'épi de graminées' (notion de botanique) ;
- l'arête, 'partie du squelette d'un poisson' (notion d'ichtyologie) ;
- l'arête, 'ligne d'intersection de deux plans' (notion de géométrie dans l'espace, d'architecture, etc.).

(6.o)
Munich [mynik] (München en allemand, Minga en bavarois) est, avec 1 443 122 habitants, la troisième ville d'Allemagne par la population après Berlin et Hambourg.

De façon orthogonale, les connaissances portées par la SE peuvent être contextualisées dans l'espace (SE 6.h, 6.m), dans le temps, ou dans tout autre dimension (SE 6.n), à l'aide de circonstants. Ceux-ci permettent d'envisager l'identification de relations autres que binaires. Nous distinguons les **SE contextuelles** (avec au moins un circonstant) des **SE non contextuelles** (sans circonstant).

6.2 Campagne d'annotation

La typologie ouvre la voie à une caractérisation plus fine des SE en vue d'un procédé d'extraction de relations à partir de celles-ci. Afin d'obtenir un retour quantitatif sur cette typologie, mais également de mettre au jour des indices attestés en corpus, nous avons utilisé la typologie pour établir un schéma d'annotation et débuter une campagne d'annotation. La tâche d'annotation s'est déroulée en trois étapes :

1. **Annotation visuelle des SE** : Cette étape visait à délimiter la SE et ses différents composants (amorce, items, clôture) lorsqu'elle bénéficiait d'une mise en forme. Cette étape est liée à l'axe visuel.

2. **Annotations rhétorique, intentionnelle et sémantique des SE** : Cette étape visait à annoter la SE selon les axes rhétoriques, intentionnels et sémantiques définis précédemment. Chaque SE se voit affecter un type rhétorique, un ou plusieurs types intentionnels et un type sémantique dénotant la relation sémantique portée.

3. **Annotation des entités textuelles dans les SE** : Lorsque le type sémantique porté par la SE est utile à la construction de ressources, cette étape visait à délimiter les segments de texte qui dénotent l'entité textuelle dans l'amorce et les entités textuelles dans chacun des items reliés par la relation sémantique annotée.

La tâche d'annotation portait initialement sur les SE horizontales avec mise en forme typographique et les SE verticales avec mise en forme typographique et dispositionnelle. Cela sera le cas dans l'annotation visuelle. Dans les deux étapes suivantes, nous avons

restreint notre travail aux SE verticales, car identifiables et manipulables avec notre modèle représentant la structure hiérarchique des documents (Chapitres 4 et 5). Les SE horizontales annotées pourront faire l'objet d'un travail ultérieur.

Les documents utilisés pour l'annotation ont été collectés en utilisant l'ontologie OntoTopo, construite dans le cadre de l'ANR GEONTO[2]. Cette ontologie modélise les domaines de l'aménagement urbain, l'environnement et l'organisation territoriale. Le corpus a été construit en projetant les étiquettes lexicales des concepts de l'ontologie OntoTopo sur Wikipédia et en récupérant les documents correspondant à ces étiquettes (p. ex. Arbre, Pont, Abbaye, etc.). Par leur caractère encyclopédique, ces documents Wikipédia ordonnent de nombreuses définitions et propriétés au moyen de marqueurs typo-dispositionnels. Au final, 169 documents furent extraits.

Dans cette section, nous décrivons l'outil d'annotation LARAt. Ensuite, nous présentons les traitements et les résultats associés aux étapes de l'annotation du corpus.

6.2.1 Outil d'annotation LARAt

Pour être menée à bien, cette tâche d'annotation nécessitait un outil adapté à la caractérisation multi-dimensionnelle des SE, cas moins courant en annotation où l'on privilégie habituellement des annotations avec des étiquettes sur un seul axe. De plus, il était également indispensable que cet outil puisse supporter le caractère imbriqué et potentiellement récursif des SE. Par exemple, une SE peut contenir d'autres SE et elle-même être imbriquée au sein d'une structure discursive plus large (p. ex. citation) ou être étalée sur plusieurs d'entre elles (p. ex. un titre et plusieurs paragraphes).

Les outils d'annotation tels que MMAX2 (Müller et Strube, 2006), MAE (Stubbs, 2011) ou encore Glozz (Widlöcher et Mathet, 2009) ne répondent pas ou partiellement à ces exigences. MMAX2 et MAE prennent du texte brut en entrée et ne gardent pas la mise en forme des textes. Glozz, conçu pour l'annotation dans l'ANR ANNODIS, supporte la mise en forme du texte mais n'est, en l'état, pas adapté pour une annotation rapide et ergonomique d'objets multi-étiquettes. En outre, la possibilité de faire évoluer le code source de Glozz n'est pas assurée (licence restrictive).

Pour toutes ces raisons, nous avons développé LARAt (*Logiciel d'Acquisition de Relations par l'Annotation de textes*[3]), prononcé /laʁa/. Le code source Java suit un paradigme Modèle-Vue-Contrôleur pour faciliter sa maintenance, et est redistribué librement. Dans son état actuel, LARAt prend en entrée des fichiers HTML ou XML respectant la norme TEI[4], les affiche en respectant leur structure logique et permet l'annotation et la caractérisation de SE qui se superposent à la structure logique.

[2] Collaboration entre le COGIT, le LRI, le LIUPPA et l'IRIT - http://geonto.lri.fr
[3] (en) *Layout Annotation for Relations Acquisition tool*
[4] Text Encoding Initiative - http://www.tei-c.org/

La figure 6.4 présente une capture d'écran de LARAt. Le panneau de gauche contient le document en cours d'annotation. Il s'agit d'une représentation logique : les éléments de mise en forme ont été préalablement balisés et leur forme visuelle suit une convention graphique prédéfinie. Le texte surligné désigne les SE précédemment annotées, ou la SE en cour d'annotation (dite courante) selon la couleur du surlignage.

Le panneau central représente la SE courante hors de son co-texte et permet la délimitation de ses composants (amorce, items, clôture) ainsi que des segments textuels participant à la relation sémantique (entités textuelles, circonstants, marqueurs de relation). Un code couleur est utilisé pour baliser chaque composant et segment textuel.

Le panneau de droite permet la caractérisation de la SE avec les types relatifs aux axes visuel, rhétorique, intentionnel et sémantique. Dans l'axe sémantique, les sous-types de *à visée ontologique* sont déroulés et signalés par les noms de leurs relations (*hyperonymie*, *instance-de*, *holonymie*, *ontologique_autre*). Il est également possible de dire si la SE est contextuelle ou non, en cas de présence d'un circonstant. Enfin, les boutons de validation et de contrôle sont situés en bas du panneau.

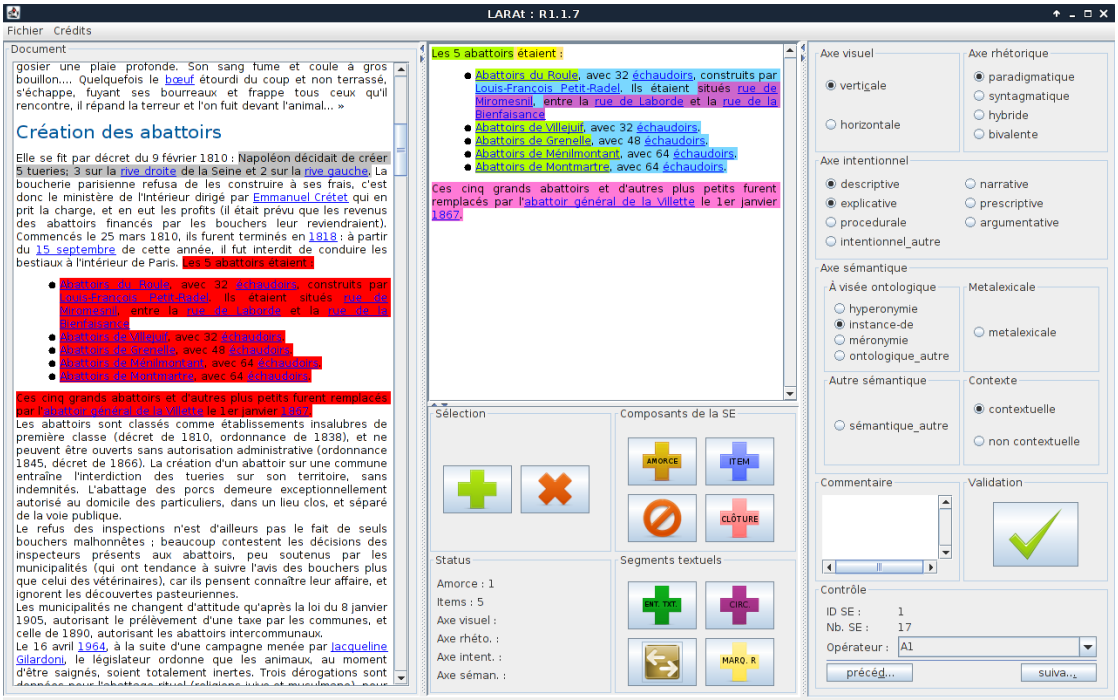

FIGURE 6.4 : Capture d'écran de l'outil d'annotation LARAt implémentant la typologie multi-dimensionnelle. Le document annoté est Abattoir.

Dans l'implémentation, deux types d'annotation sont produits (type 1 et type 2). Les annotations de type 1 concernent exclusivement l'annotation visuelle des SE (Figure 6.5). Une fois délimitée, les SE sont caractérisées avec des annotations de type 2 (Figure 6.6). Celles-ci concernent les annotations rhétorique, intentionnelle et sémantique, ainsi que l'annotation des entités textuelles. Cette manière modulaire de gérer l'annotation facilite les post-traitements (p. ex. étude d'un phénomène particulier, recherche d'un cas précis pour exemplifier un emploi, etc.).

```
<annotation type="1">
  <metadata>
    <id>1</id>
    <auteur>A1</auteur>
    <date>13/07/13 10:54</date>
    <document>Abattoir.html</document>
  </metadata>
  <SE>
    <id>1</id>
    <text>Les 5 abattoirs étaient :
    Abattoirs du Roule, (...)
          ...
    (...) de la Villette le 1er janvier 1867.</text>
    <span begin="2606" end="3110" />
  </SE>
  <comment />
</annotation>
```

FIGURE 6.5 : Exemple du format XML des annotations de type 1

6.2.2 Annotation visuelle des SE

La délimitation des SE et de leurs composants a constitué la première étape de l'annotation. Deux annotateurs étudiants ont participé à cette tâche. Rappelons que celle-ci portait sur les SE horizontales mises en forme typographiquement, ainsi que sur les SE verticales mises en forme typographiquement et dispositionnellement.

Alignement visuel des SE Afin de calculer l'accord entre les annotateurs et de travailler sur un sous-ensemble commun dans les phases d'annotation suivantes, nous avons procédé à un alignement positionnel des SE annotées. Un alignement est un lien fait entre les unités annotées par deux ou plusieurs annotateurs. Dans notre cas, chaque alignement entre deux unités est représenté par les différences absolues entre leur index de début et leur index de fin. Au plus ces différences sont proches de zéro, au plus il est probable que les annotateurs aient annoté le *même* phénomène. Pour l'ensemble du corpus, les deux annotateurs ont délimité respectivement 1406 et 1517 segments textuels comme étant des SE, et 31496 configurations d'alignements étaient possibles.

```
<annotation type="2">
  <metadata>
    <id>1</id>
    <auteur>A1</auteur>
    <date>13/07/13 10:54</date>
    <document>Abattoir.html</document>
  </metadata>
  <SEval type="axe_semantique" idSE="1">
    <visee_ontologique value="1">
      <hyperonymie value="0" />
      <instancede value="1" />
      <meronymie value="0" />
      <ontologique_autre value="0" />
    </visee_ontologique>
    <metalinguistique value="0" />
    <semantique_autre value="0" />
    <contextuelle value="1" />
  </SEval>
</annotation>
```

FIGURE 6.6 : Exemple du format XML des annotations de type 2

Pour obtenir une configuration d'alignements idéale, nous avons réalisé l'alignement en deux étapes. Premièrement, nous avons implémenté et utilisé l'algorithme de Mathet et Widlöcher (2011). Il s'agit d'un algorithme permettant d'obtenir une solution approchée d'une configuration d'alignements idéale (Annexe C.1). Deuxièmement, nous avons vérifié et nettoyé manuellement les alignements avec une interface. En figure 6.7, nous donnons un exemple de cette interface. Chaque ligne horizontale correspond au fil du texte pour un des annotateurs et les rectangles correspondent aux segments délimités comme SE. Les liens entre les rectangles sont les alignements entre les SE annotées. En annexe C.2, nous donnons des exemples supplémentaires avec cette interface.

Accord pour l'annotation visuelle des SE Pour mesurer l'accord entre les deux annotateurs, nous utilisé une méthode identique à celle des campagnes MUC (Chinchor et Marsh, 1998). Il s'agit de calculer une mesure de F_1-score en considérant une des annotations comme celle de référence[5]. Dans ce cadre, le F_1-score est défini comme :

$$F_1 - score = 2 * \frac{|X \cap Y|}{|X| + |Y|} \qquad \text{(eq.6.1)}$$

où X est l'ensemble des segments textuels considérés comme SE par le premier annotateur, et Y est l'ensemble des segments textuels considérés comme SE par le second.

[5] Comme la mesure de F_1-score est symétrique, le choix de l'annotation considérée comme référence n'a pas de conséquence sur le score obtenu.

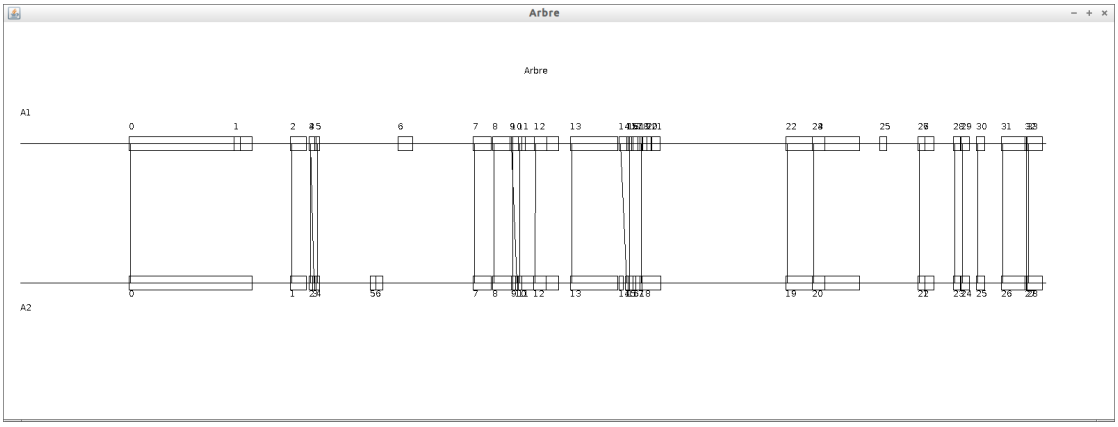

FIGURE 6.7 : Interface pour la vérification et la correction des alignements des SE. Le document traité est Arbre.

L'intersection $|X \cap Y|$ correspond à l'alignement effectué plus haut. Pour notre tâche d'annotation visuelle, le F_1-score obtenu est de 83,21. Ce score dénote une certaine stabilité des indices typo-dispositionnels pour le marquage visuel des SE.

En annexe C.3, nous répertorions pour chaque document les annotations apportées par les deux annotateurs.

Résultats de l'annotation Au terme de cette première étape, 472 SE horizontales et 745 SE verticales ont été annotées et alignées au sein de 169 documents. Une fois les annotateurs en accord sur la présence d'une SE, la délimitation de l'amorce et des items ne montrait pas de désaccord. Le tableau 6.1 présente les caractéristiques de l'ensemble des 1217 SE. Les SE verticales montrent en moyenne un plus grand nombre d'items que les SE horizontales. La SE verticale présentant 84 items est une SE listant les communes françaises dont le toponyme est lié lexicalement au terme fontaine dans le document du même nom. La SE horizontale présentant 16 items est une SE énumérant des animaux dans le document Cirque (SE 6.p).

	Nombre SE	Nombre items	Min-Max items	Moyenne par SE
SE verticales	745	4145	2 - 84	5,56
SE horizontales	472	1771	2 - 16	3,75
Corpus	1217	5916	2 - 84	4,86

TABLE 6.1 : Caractéristiques des SE délimitées et alignées par les deux annotateurs dans l'ensemble du corpus

(6.p)	Dressage et domptage d'animaux : autruche, chameau, cheval, chien, dromadaire, éléphant, girafe, lama, lion, otarie, ours, panthère, serpent, singe, tigre, etc.

6.2.3 Annotations rhétorique, intentionnelle et sémantique des SE

Pour cette phase d'annotation, les deux annotateurs étudiants ont procédé à l'annotation des axes rhétorique, intentionnel et sémantique des SE délimitées et alignées. Au vu de la difficulté de ces tâches, un annotateur considéré expert a été ajouté afin d'arbitrer les cas problématiques.

Axe rhétorique Dans l'axe rhétorique, les annotateurs avaient le choix entre les types mutuellement exclusifs *paradigmatique*, *syntagmatique*, *hybride* et *bivalente*. Pour mesurer l'accord sur ces annotations catégorielles nous utilisons un κ de Fleiss (1979). Celui-ci est une généralisation du π de Scott (1955) à plus de deux annotateurs. Il s'agit essentiellement de calculer l'accord en le corrigeant par une mesure de hasard :

$$\kappa = \frac{A_o - A_a}{1 - A_a}$$

où A_o est l'accord observé dans les données, et A_a une estimation du hasard. La valeur obtenue est une valeur comprise entre 0 et 1 : un score bas indique un désaccord (systématique à 0) et un score haut indique un accord (systématique à 1) entre les annotateurs[6].

Nous avons obtenu un κ de 0,21. Ce score est considéré comme faible sur l'échelle de Green (1997). Ce résultat montre que les annotateurs ont eu des difficultés avec les notions rhétoriques[7]. Considérant cette instabilité ainsi que la difficulté d'effectuer un alignement sur cet axe, nous nous sommes référés uniquement à l'annotation effectuée par l'expert. Dans ce contexte, il apparaît que le type rhétorique *Paradigmatique* est associé significativement (χ^2 avec un risque α à 0,05) à la présence d'une relation à visée ontologique ou metalexicale. Ceci semble confirmer la tendance des SE paradigmatiques verticales à être porteuses de relations utiles à la construction de ressources. Toutefois, une nouvelle phase d'annotation avec des annotateurs mieux formés sur cet axe serait nécessaire pour confirmer ce résultat.

Axe intentionnel L'axe intentionnel présentait un cas d'annotation multi-étiquettes. Les annotateurs avaient pour tâche d'annoter les SE selon les types *descriptive*, *narrative*, *explicative*, *prescriptive*, *procédurale*, *argumentative*, *intentionnel_autre*.

[6] L'intuition derrière la mesure de hasard A_a est que si l'accord observé A_o est élevé, mais que A_a l'est aussi, alors c'est qu'il est probable que l'accord ait été obtenu par hasard. Dans ce cas, le score final tendra vers 0. Les principales mesures d'accord (S de Bennett *et al.* (1954), κ de Cohen (1960), etc.) se différencient par leur estimation du hasard (Artstein et Poesio, 2008).

[7] En particulier, un annotateur étudiant a annoté plus d'un tiers de ses SE comme syntagmatiques. Une erreur de compréhension en est certainement la cause.

Le caractère non-mutuellement exclusif de ces types rend difficile une mesure d'accord pour cet axe. Rosenberg et Binkowski (2004) ont proposé une version modifiée du κ pour ce genre de cas. Toutefois, si leur mesure est théoriquement intéressante, sur le plan pratique elle tend à sous-estimer l'accord observé et elle est peu répandue.

Nous avons préféré considérer chaque type intentionnel isolément afin de calculer son lien avec l'axe sémantique. Ces opérations sont alors répétées pour chaque annotateur. Les résultats montrent que, dans la très grande majorité des SE verticales, les annotateurs ont annoté le type *Descriptive*. Pour les trois annotateurs, ce type intentionnel est significativement associé à la présence d'une relation à visée ontologique ou metalexicale (χ^2 avec un risque α à 0,05). Cette prévalence descriptive s'explique en partie par un biais induit par la nature du corpus, ici encyclopédique. Les résultats montrent également que les types *Argumentatif* et *Intentionnel_autre* ne sont pas associés à la présence d'une relation sémantique. Enfin, concernant les autres types intentionnels, il est difficile d'établir des conclusions à cause des variations entre les trois annotateurs. Une nouvelle phase d'annotation avec une simplification des étiquettes doit être envisagée.

Axe sémantique Dans l'axe sémantique, les annotateurs avaient pour tâche d'associer un type sémantique aux SE. Les types proposés étaient *hyperonymie*, *instance-de*, *holonymie*, *ontologique_autre*, *metalexicale* et *sémantique_autre*.

La position centrale de cet axe dans notre typologie multi-dimensionnelle, nous a amenés à mesurer l'accord de manière systématique pour chacun des types. Pour cela, nous avons utilisé un κ de Fleiss (1979) (défini précédemment). Les résultats sont reportés dans la table 6.2. Les Z-scores montrent que les scores obtenus sont significativement meilleurs qu'une distribution aléatoire (α à 0,05). Pour l'ensemble du corpus, nous obtenons un κ 0,49. Ce score est considéré comme moyen sur l'échelle de Green (1997).

Types sémantiques	Kappa κ	Z-score
hyperonymie	0,45	18,27
instance-de	0,43	17,40
holonymie	0,48	19,53
ontologique_autre	0,28	11,39
metalexical	0,74	30,40
sémantique_autre	0,23	09,50
Corpus	0,49	36,20

TABLE 6.2 : Accords inter-annotateurs par types sémantiques

Deux confusions apparaissent dans les annotations :

- La première concerne les types sémantiques *ontologique_autre* et *sémantique_autre* qui présentent les scores d'accord les plus bas. Ceci semble s'expliquer par le manque d'indices discriminants et la rareté des observations compliquant l'établissement de régularités.

- La seconde confusion apparaît entre les types *hyperonymie* et *instance-de*. Autrement dit, il s'agit de la distinction entre les entités textuelles qui réfèrent à une classe et celles qui réfèrent à une instance. Si cette distinction est classiquement faite en représentation des connaissances (Woods, 1975) (en particulier dans le domaine des ontologies (Gangemi *et al.*, 2001)), elle n'est pas toujours simple à appliquer dans une visée d'analyse de textes (Chapitre 1). Lorsque les SE porteuses de ces deux types sémantiques sont fusionnées, nous obtenons un κ de 0,54.

Inversement, le type sémantique *metalexical* apparaît être relativement bien défini. Ceci pourrait s'expliquer par un biais du corpus : dans de nombreux cas, des SE sont utilisées pour définir, différencier ou traduire des unités lexicales dans Wikipédia.

Afin de permettre une utilisation empirique de ce corpus, nous avons aligné les SE au niveau de l'axe sémantique. Cet alignement a été effectué de manière semi-automatique. Lorsqu'au moins deux annotateurs apparaissent être en accord sur un type sémantique, ce dernier est retenu dans le corpus final. Seuls un sous-ensemble de SE ont nécessité l'intervention de l'expert pour choisir le type de référence. Ces SE sont notamment des cas ambigus ou des cas d'abstention de réponse de la part d'un des annotateurs. La distribution des SE alignées selon l'axe sémantique est présentée dans la table 6.3.

Types sémantiques	Nombre SE	Couverture %
hyperonymie	268	36,0
instance-de	196	26,3
holonymie	39	5,2
ontologique_autre	42	5,7
metalexicale	149	20,0
sémantique_autre	51	6,8
Corpus	745	100

TABLE 6.3 : Distribution des SE alignées par types sémantiques

6.2.4 Annotation des entités textuelles dans les SE

Une fois la SE annotée selon les axes rhétorique, intentionnel et sémantique, les annotateurs ont dû délimiter, lorsqu'elles étaient présentes, les entités textuelles participant à la relation. Ces entités textuelles peuvent être des termes ou des entités nommées[8].

Accord pour la délimitation des entités textuelles Afin de calculer l'accord entre les trois annotateurs, nous avons utilisé une méthode identique à celle suivie pour l'accord de la délimitation des SE (Section 6.2.2). Il s'agit d'utiliser une mesure de F_1-score en considérant une des annotations comme celle de référence.

[8] Nous avons initialement fait l'hypothèse que le choix du type sémantique par les annotateurs permettrait de distinguer les termes et les entités nommées.

Dans notre cas à trois annotateurs, nous avons évalué toutes les paires d'annotateurs tour à tour et nous avons moyenné[9] les résultats en distinguant les amorces et les items. La table 6.4 présente les résultats obtenus. Les scores pour les items sont systématiquement supérieurs à ceux obtenus pour les amorces. Ceci semble indiquer que les entités textuelles impliquées dans la relation et présentes dans les items montrent des caractéristiques d'apparition plus stables que celles présentes dans l'amorce.

Annotateurs	Amorces	Items	Moyenne
Expert vs. Étudiant 1	68,31	70,32	69,31
Expert vs. Étudiant 2	70,46	88,77	79,61
Étudiant 1 vs. Étudiant 2	70,05	75,34	72,70
Moyenne	69,60	78,14	73,87

TABLE 6.4 : Mesures de F_1-score pour l'accord sur la délimitation des entités textuelles

Coordination L'annotation a fait émerger un cas particulier : la présence de plusieurs entités textuelles impliquées dans la relation et situées dans une même amorce ou dans un même item. Généralement, ces entités textuelles apparaissent coordonnées syntaxiquement et peuvent elles-mêmes former une énumération horizontale. La SE (6.q), tirée du document Culte, est intéressante sur ce point. Celle-ci est porteuse d'une relation d'hyperonymie où l'hyperonyme est *actes cultuels* et où les hyponymes forment dans plusieurs items des groupes coordonnés syntaxiquement.

(6.q)

Les principaux actes cultuels sont :
- le sacrifice, la libation, l'offrande et l'éducation ;
- la prière (invocation, louange, demande, etc.) ;
- le chant et la musique ;
- la lecture de textes sacrés le cas échéant ;
- éventuellement la prédication qui a un rôle important surtout dans les religions abrahamiques et le bouddhisme (mais la prédication peut aussi s'effectuer dans le cadre d'une activité missionnaire qui n'est pas liée à un culte proprement dit) ;
- les pèlerinages, processions.

6.3 Discussion

L'analyse linguistique que nous avons menée sur les SE a permis de définir une typologie multi-dimensionnelle, permettant de tenir compte de propriétés de natures différentes et parfois orthogonales. Sur le plan théorique, ce travail a pu éclairer en partie le phénomène

[9] Cette manière de procéder est reprise de Tateisi *et al.* (2000).

complexe des SE quant à leur forme, leur structure et leur fonction. Une tâche d'annotation a été menée en corpus. À cet égard, nous avons développé l'outil d'annotation LARAt qui permet de catégoriser les SE suivant les différents axes de notre typologie. Enfin, nous avons apporté un regard quantitatif sur les résultats obtenus au travers de mesures statistiques.

Le choix du corpus induit des biais dont il faut être conscient. Dans notre cas, deux biais peuvent être soulevés. Premièrement, les conventions de rédaction de Wikipédia[10] peuvent expliquer le nombre relativement élevé de SE paradigmatiques. Ces conventions préconisent une forme grammaticale identique pour tous les items des SE. Deuxièmement, le caractère encyclopédique de Wikipédia induit que les SE sont pour, dans leur grande majorité, descriptives. Ainsi, il n'est pas surprenant que les SE annotées comme narratives, prescriptives ou procédurales soient moins fréquentes.

L'annotation est un processus difficile, car celle-ci implique l'interprétation. La manière de guider celle-ci est complexe et dépend des besoins. Un équilibre doit être trouvé entre une tâche où les annotateurs sont considérés comme indépendants[11] afin d'évaluer un schéma d'annotation (Krippendorff, 1980) et une tâche où les annotateurs sont suffisamment contraints que pour avoir des données utilisables dans un procédé d'apprentissage (Pustejovsky et Stubbs, 2012). Parallèlement, le risque qu'un annotateur n'ait pas compris la tâche d'annotation[12] est d'autant plus élevé que la tâche est complexe.

Dans ce contexte, il est important de considérer le *rôle exploratoire* qui a été donné à l'annotation dans ce chapitre. Ceci apparait notamment dans les accords bas sur certains axes ou dans le caractère difficilement exploitable de certaines données.

Pour une utilisation pratique des données annotées, il a été nécessaire de simplifier et de nettoyer celles-ci. Nous avons effectué plusieurs post-traitements (normalisation, élimination des doublons, etc.) de manière incrémentale jusqu'à obtenir un jeu de données consistant sur lequel nous avons pu construire et évaluer notre méthode d'extraction de relations dans les SE (Chapitre 7).

Notons que le corpus et l'outil LARAt associé sont librement accessibles[13] et modifiables conformément à la licence Creative Commons BY-NC-SA 3.0[14] pour le corpus, et la licence CeCILL 2.1[15] pour l'outil LARAt.

[10] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Conventions_typographiques

[11] Les mesures d'accord catégorielles (S de Bennett *et al.* (1954), π de Scott (1955), etc.) font l'hypothèse d'indépendance des événements (d'annotation) : $p(A \cap B) = p(A).p(B)$

[12] Le cas du document Glacier est particulièrement parlant (Annexe C.3).

[13] Corpus - https://github.com/fauconnier/corpus-LARA
Outil LARAt - https://github.com/fauconnier/LARAt

[14] https://creativecommons.org/licenses/by-nc-sa/3.0

[15] http://www.cecill.info/licences/Licence_CeCILL_V2.1-fr.txt

Chapitre 7

Extraction de relations sémantiques dans les structures énumératives paradigmatiques verticales

Sommaire

Dans ce chapitre nous décrivons une méthode pour extraire les relations sémantiques à partir de SE paradigmatiques verticales. Cette méthode s'appuie sur la représentation de la structure de document décrite dans la partie II, et sur le travail de caractérisation linguistique des SE (Chapitre 6). L'extraction d'une relation est effectuée à l'aide de trois tâches :

1. **Identification des SE d'intérêt** : Nous ciblons des SE présentant des propriétés rhétoriques et visuelles distinctes. Celles-ci sont appelées SE d'intérêt. Cette étape propose de les identifier dans l'arbre de dépendances représentant la structure logique. Cette identification s'effectue en parcourant l'arbre à la recherche de motifs.

2. **Qualification de la nature de la relation sémantique** : En fonction de la relation sémantique recherchées, nous spécifions si les SE d'intérêt sont porteuses ou non de cette relation sémantique. Cette qualification s'effectue selon des critères typographiques et lexico-syntaxiques.

3. **Identification des arguments de la relation sémantique** : Si la relation sémantique portée par une SE d'intérêt est une relation recherchée, les entités textuelles (Section 1.1.3) constituant les arguments de la relation présents dans la SE sont identifiées. Cette identification s'effectue selon des critères typographiques, dispositionnels, lexicaux et syntaxiques.

La première étape dépend de la bonne construction de la structure logique du document en amont. Les deux tâches suivantes reposent sur de l'apprentissage supervisé et nécessitent préalablement des données d'entraînement. C'est pourquoi nous avons utilisé le corpus annoté selon notre typologie décrit en section 6.2. La figure 7.1 schématise l'ensemble du procédé.

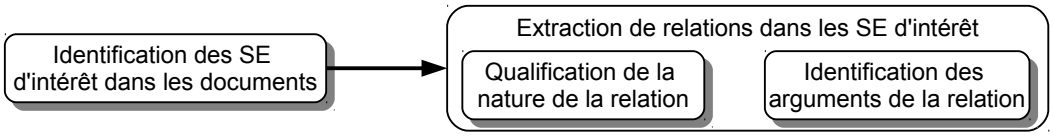

FIGURE 7.1 : Schéma du système pour l'extraction de relations sémantiques dans les structures énumératives d'intérêt

Dans ce chapitre, nous étendons la définition de l'hyperonymie pour prendre en compte la relation « instance de ». L'hyperonymie couvre des phénomènes qui peuvent donner lieu à deux types de représentations sémantiques : (i) une classe appartenant au sous-ensemble d'une autre classe, et (ii) un individu appartenant à une classe. Or la distinction entre ces deux situations constitue une tâche complexe à part entière, qu'il est possible de réaliser en aval de notre travail.

Par conséquent, le passage entre la modélisation linguistique et la modélisation conceptuelle n'est pas effectué (Chapitre 1) et nous ne faisons pas de distinction entre les relations entre classes (p. ex. *université* et *université nationale*), et celles impliquant une instance (p. ex. *université* et *université de Toulouse*). Dans ce contexte, notre position se rapproche notamment de celles de Miller *et al.* (1990)[1], Hearst (1992)[2], Snow *et al.* (2004) ou encore Sumida et Torisawa (2008).

La suite de ce chapitre présente les trois tâches pour l'extraction des relations sémantiques. Ensuite, nous proposons d'évaluer la précision de l'ensemble du système sur des données non-annotées.

[1] Dans le cas de WordNet, il faudra attendre Miller et Hristea (2006) pour qu'une première distinction entre classes et instances soit faite. Les auteurs expliquent : « (...) in some cases the distinction was not easy to draw. Incorporating this distinction was resisted at first because WN was not initially conceived as an ontology, but rather as a description of lexical knowledge. »

[2] Voir notamment l'exemple (3.f) à la page 90.

7.1 Identification des structures énumératives d'intérêt

Dans cette section, nous décrivons les SE recherchées, appelées SE d'intérêt, et le parcours d'arbre visant leur identification dans la structure logique du document.

7.1.1 Description

Nous proposons une méthode pour l'identification des SE intéressantes pour l'extraction de relations, dans la structure logique du document. Nous considérons cette tâche comme un problème de filtrage par motifs : l'hypothèse est faite que l'énumération d'items est plus simple à identifier et peut conduire au repérage de l'amorce. Deux points sont discutés :

- la forme des structures énumératives d'intérêt ;
- la recherche de celles-ci dans le document.

Structures énumératives d'intérêt Dans ce travail, nous avons choisi d'exploiter les structures énumératives qui présentent deux propriétés : (i) sur le plan rhétorique, elles sont paradigmatiques et (ii) sur le plan visuel, elles sont marquées typographiquement et dispositionnellement. Des liens peuvent être faits respectivement avec la typologie de Luc (2000) et les SE de Type 1 (sections titrées) et de Type 2 (listes formatées) dans la typologie de Ho-Dac *et al.* (2010).

Nous faisons l'hypothèse que si une SE présente une coénumérabilité visuelle, c'est-à-dire que son énumération est homogène, alors elle est paradigmatique. Cette simplification permet de définir les SE d'intérêt comme celles qui rencontrent les deux propriétés suivantes dans la structure logique du document :

1. Ces structures énumératives présentent m unités logiques élémentaires coordonnées entre elles. Celles-ci sont les items de la structure énumérative.

2. La première des unités logiques élémentaires coordonnées est subordonnée à une autre unité logique élémentaire, appelée amorce, qui introduit l'énumération.

La figure 7.2 donne la représentation d'une telle structure selon le modèle introduit dans le chapitre 4. Pour rappel, la relation de subordination est représentée par une flèche pleine et la relation de coordination est représentée par une flèche en pointillé.

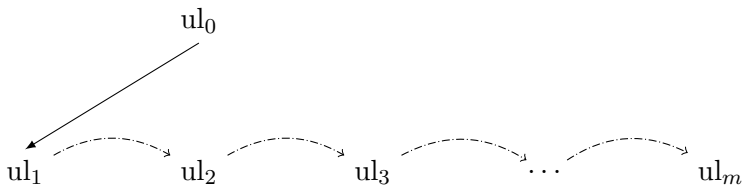

FIGURE 7.2 : Représentation en dépendances de la forme des SE d'intérêt

Bien qu'il y ait une seule relation de subordination explicite (entre ul_0 et ul_1), ce type de structure implique $m - 1$ relations de subordination implicites (entre entre ul_0 et ul_i $(1 < i \leq m)$). Rappelons que, dans notre modèle de représentation de la structure logique, si une ul_j est subordonnée à une ul_i, alors toutes les ul_k coordonnées à ul_j sont également subordonnées à ul_i.

Dans ce cadre, si la SE est porteuse d'une relation sémantique, il est idéalement attendu que celle-ci soit distribuée entre ul_0 et ul_i $(1 \leq i \leq m)$. Cela se traduit par une entité textuelle portée par l'amorce et des entités textuelles, dans les items, liées à celle-ci selon un même type de relation sémantique.

Recherche de SE d'intérêt dans le document L'identification des SE d'intérêt s'effectue au moyen d'un filtrage par motifs. Les motifs décrivent un ensemble de contraintes auxquelles doivent répondre les structures textuelles pour pouvoir être identifiées comme SE d'intérêt. Les contraintes portent notamment sur la forme de la structure qui doit être comparable à celle décrite en figure 7.2. Additionnellement, le nombre d'unités logiques élémentaires coordonnées doit être supérieur à deux. D'autres contraintes portent sur la nature des étiquettes des unités logiques : les unités logiques coordonnées doivent porter une étiquette d'item, tandis que l'étiquette logique portée par l'amorce peut être un titre, un paragraphe ou un autre item.

Dans la pratique, l'identification suit la procédure suivante : (i) lorsqu'une unité logique étiquetée comme item est rencontrée, (ii) l'amorce est recherchée dans l'unité logique qui la subordonne directement, ensuite (iii) le reste de l'énumération est cherché dans les unités logiques dépendantes (coordonnées et subordonnées) à la première. Dans ce contexte, les items peuvent eux-mêmes constituer de sous-arbres.

L'exemple (7.a), extrait de notre corpus, est segmenté en unités logiques élémentaires. L'arbre de dépendances correspondant à cet extrait est donné en figure 7.3. Quatre SE d'intérêt sont identifiées selon les contraintes citées ci-dessus :

- La SE constituée des unités logiques élémentaires (ul_1, ul_2, ul_3, ..., ul_{14}),
- La SE constituée des unités logiques élémentaires (ul_4, ul_5, ul_6, ul_7, ul_8),
- La SE constituée des unités logiques élémentaires (ul_9, ul_{10}, ul_{11}, ul_{12}, ul_{13}),
- La SE constituée des unités logiques élémentaires (ul_{11}, ul_{12}, ul_{13}).

Les SE horizontales au sein des unités logiques (p. ex. celle dans ul_8) ne sont pas prises en compte ici.

La méthode proposée fonctionne à condition que la construction de l'arbre de dépendances soit correctement faite en amont. Si cela n'est pas systématique pour le format PDF (Chapitre 5), c'est néanmoins majoritairement le cas pour les formats à balises, pour lesquels le problème est considéré de manière déterministe. Notons que notre méthode est pour l'instant *ad hoc* et vise uniquement les SE répondant à la combinaison des motifs énoncés ci-dessus. Une perspective pourrait viser l'adaptation aux arbres de dépendances de l'outil Tregex (Levy et Andrew, 2006), dédié à la recherche de sous-arbres dans les arbres de constituants.

(7.a)

[**Les différents types d'aquaculture**]__1

- [L'aquaponie, polyculture extensive intégrant sous forme de symbiose poissons, mollusques, et une multiplicité de végétaux, lesquels se nourrissent des déjections elles mêmes transformées par des bactéries ;]__**2**
- [La pisciculture, c'est-à-dire l'élevage de poissons ;]__**3**
- [La conchyliculture, l'élevage de coquillages. Les types les plus courants de conchyliculture sont :]__**4**
  - [l'ostréiculture (élevage des huîtres),]__**5**
  - [l'halioticulture (élevage des ormeaux),]__**6**
  - [la mytiliculture (élevage des moules),]__**7**
  - [la pectiniculture (élevage de coquilles Saint-Jacques ou de pétoncles) ;]__**8**
- [l'élevage de crustacés :]__**9**
  - [L'astaciculture est l'élevage des écrevisses,]__**10**
  - [La pénéiculture (élevage de crevettes de mer et de crevettes d'eau douce) est pratiquée en France,]__**11**
    - [les crevettes "gambas" sont élevées en grande quantité au Brésil,]__**12**
    - [la crevette impériale ;]__**13**
- [L'algoculture, c'est-à-dire la culture d'algues.]__**14**

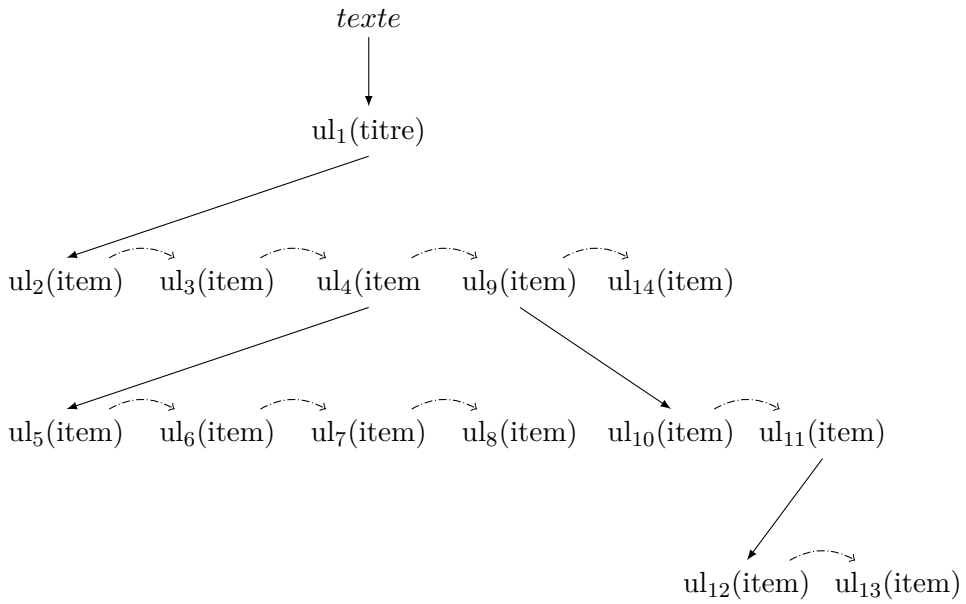

FIGURE 7.3 : Arbre de dépendances correspondant à l'exemple (7.a)

7.2 Qualification de la relation sémantique

Dans cette section, nous décrivons la méthode proposée et, ensuite, nous reportons l'évaluation obtenue sur le corpus annoté.

7.2.1 Description

Nous décrivons ici une méthode pour la qualification de la relation sémantique portée par les SE d'intérêt. Cette tâche est considérée comme un problème d'apprentissage en cascade. Dans ce cadre, plusieurs classifieurs sont utilisés en série[3] : les sorties d'un classifieur donné sont utilisées comme informations complémentaires dans le classifieur qui lui succède. Ce type d'architecture permet la décomposition d'un problème en sous-problèmes. L'une des premières applications fut le système de détection de visage proposé par Viola et Jones (2001) dans le domaine de la vision par ordinateur.

Afin d'évaluer notre méthode, nous nous sommes focalisés sur la relation d'hyperonymie, car sa fréquence d'apparition est généralement indépendante du domaine du corpus. Dans ce contexte, elle est utilisée pour définir les notions de ce domaine (Morin, 1999). Deux tâches de classification en série sont proposées :

- **T_Onto** Cette tâche cherche à identifier les SE porteuses d'une relation sémantique de type *à visée ontologique* (c'est-à-dire une relation appartenant à un des sous-types sémantiques *hyperonymie*, *holonymie* et *ontologique_autre*).
- **T_Hypo_1** Cette tâche cherche à identifier les SE porteuses de la relation d'hyperonymie (c'est-à-dire appartenant au sous-type sémantique *hyperonymie*), en utilisant additionnellement les sorties de T_Onto.

Nous comparons l'enchaînement de ces deux tâches à une tâche de classification isolée :

- **T_Hypo_2** Cette tâche cherche à identifier les SE porteuses de la relation d'hyperonymie (c'est-à-dire appartenant au sous-type sémantique *hyperonymie*), sans utiliser les sorties de la tâche T_Onto.

La figure 7.4 schématise l'ensemble du système.

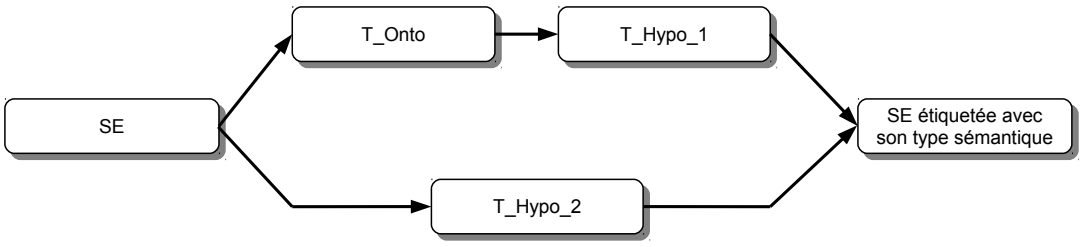

FIGURE 7.4 : Schéma du système pour l'identification de l'hyperonymie

[3] D'autres techniques permettent la combinaison de classifieurs, tels que le bagging ou le boosting, mais en utilisant les classifieurs en parallèle.

À titre d'exemple, nous cherchons à faire la distinction, dans l'exemple (7.a), entre les SE (ul_1, ul_2, ul_3, ..., ul_{14}), (ul_4, ul_5, ul_6, ul_7, ul_8) et (ul_9, ul_{10}, ul_{11}, ul_{12}, ul_{13}), porteuses d'une relation d'hyperonymie, et la SE (ul_{11}, ul_{12}, ul_{13}) qui n'en porte pas.

Notons que, contrairement au système de Viola et Jones (2001), notre système ne retranche pas les observations à chaque étape de classification. Ce choix est notamment fait à cause du nombre restreint de données à notre disposition.

Dans le reste de cette section, deux points sont abordés :
- la formalisation des tâches en un problème d'apprentissage ;
- les traits permettant de capturer les indices de la relation d'hyperonymie.

Apprentissage supervisé Pour chaque tâche de classification, nous utilisons des algorithmes d'apprentissage supervisé pour déterminer si une SE donnée porte le type sémantique ciblé par la tâche. Chaque SE est représentée par un vecteur de traits $\mathbf{x}$, capturant ses informations typographiques, lexicales et syntaxiques. L'algorithme d'apprentissage doit associer ce vecteur de traits à une classe y. Cette classe est soit positive, — la SE porte le type sémantique ciblé —, soit négative. Cette fonction cible a la forme :

$$f(\mathbf{x}) = y$$

Pour approximer f, nous utilisons et comparons deux algorithmes d'apprentissage. Le premier apprend des modèles linéaires qui souvent généralisent bien, mais qui restent sensibles aux distributions de données présentant une grande dispersion. Le second algorithme apprend des modèles non-linéaires pouvant conduire à une séparation précise des données d'entraînement, mais avec un risque accru de sur-apprentissage.

Nous avons utilisé respectivement une régression logistique (Cox, 1959) et une Machine à Vecteurs de Support (Cortes et Vapnik, 1995) avec un noyau gaussien. Nous donnons ci-dessous une brève description de ces deux algorithmes d'apprentissage.

- La régression logistique s'inscrit dans la tradition statistique de la régression linéaire (Hastie *et al.*, 2009), mais applique une fonction d'activation au produit scalaire du vecteur de traits et du vecteur de paramètres pour déterminer, de manière probabiliste, la classe du type sémantique portée par une SE. Généralement, la fonction d'activation utilisée est une fonction sigmoïde :

$$p(y|\mathbf{x}) = \frac{1}{1 + \exp(-\theta^{\mathrm{T}}\mathbf{x})} \qquad \text{(eq.7.1)}$$

  L'estimation des paramètres s'effectue en optimisant la log-vraisemblance sur les données d'apprentissage. Cette fonction étant convexe, une grande variété de processus itératifs permettent de trouver son optimum, tels que l'algorithme BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). Dans notre travail, nous utilisons l'algorithme GIS (*Generalized Iterative Scaling*) proposé par Darroch et Ratcliff (1972).

- Les Machines à Vecteurs de Support, ci-après SVM, s'inscrivent dans la tradition des algorithmes à hyperplan séparateur (Rosenblatt, 1958), mais en proposant deux principes supplémentaires : (i) la maximisation des marges autour de l'hyperplan permet d'améliorer la généralisation du modèle (Vapnik, 1995) et (ii) l'utilisation de fonctions noyaux permet de transformer l'espace de traits en un espace de plus grande dimension afin d'apprendre des fonctions non linéaires (Aizerman *et al.*, 1964). Ces deux principes sont rassemblés dans la forme du SVM proposée par Boser *et al.* (1992). Dans ce cadre, les vecteurs de support sont des exemples appris par lesquels passent les marges. La fonction de décision a la forme :

$$f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{v} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b) \qquad \text{(eq.7.2)}$$

où b est le biais, y_i et α_i sont respectivement la classe et le multiplicateur de Lagrange associés au i-ème vecteur de support. Dans ce travail, nous utilisons une fonction gaussienne $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2)$. Celle-ci peut être considérée comme une fonction de similarité entre la SE à classer $\mathbf{x}$ et le vecteur de support $\mathbf{x}'$. Le γ définit la sensibilité de cette similarité.
L'apprentissage vise à choisir les vecteurs de support et les multiplicateurs de Lagrange associés qui maximisent les marges. Il s'agit d'un problème d'optimisation quadratique sous contraintes, qu'il est possible de résoudre avec des techniques classiques d'optimisation (Goldfarb et Idnani, 1982). Toutefois, dans notre travail, nous utilisons une technique dédiée au SVM, la décomposition SMO (*Sequential Minimal Optimization*) proposée par Platt *et al.* (1998).

Des explications plus complètes ainsi qu'une comparaison entre les algorithmes peuvent être trouvées dans l'annexe consacrée à l'apprentissage supervisé (Annexe B).

Traits utilisés Nous décrivons ici les traits utilisés pour capturer les informations typographiques, lexicales et syntaxiques des SE. L'ensemble de traits est identique pour les trois tâches de classification. Seule la tâche T_Hypo_1 utilise additionnellement les sorties de T_Onto. Le choix des traits a été fait sur la base d'observations en corpus, dans la partie d'analyse linguistique de notre travail (Chapitre 6).

Les informations capturées prennent en compte uniquement l'amorce de la SE et son premier item. Le choix de ne pas prendre des informations à propos de tous les items a été justifié dans un travail préliminaire (Fauconnier *et al.*, 2013b) : les résultats suggèrent que l'identification de la relation sémantique en utilisant l'amorce et le premier item est à ce jour la meilleure approche[4], surtout lorsque les SE présentent un grand nombre d'items. Notons que ce travail a également montré la supériorité des traits conçus à la main (*handcrafted features*) par rapport à des traits appris (*learned features*) tels que les trigrammes de tokens.

[4] Ce travail antérieur a comparé deux configurations : (i) une configuration qui prédit la première paire amorce-item, (ii) une configuration qui fait la moyenne des prédictions sur toutes les paires amorce-item d'une SE donnée. Se référer à l'article associé pour les détails (Fauconnier *et al.*, 2013b).

Les traits sont séparés en deux familles (Tableau 7.1) :

- **f_amorce_item** Cette famille de traits s'applique à l'amorce et à l'item. Les traits de type *t_POS_c* et *t_POS_p* captent des régularités au travers des catégories morpho-syntaxiques telles que la présence d'un nom commun pluriel dans l'amorce, ou l'utilisation d'un infinitif en début d'item. Les traits *t_NbToken* et *t_NbSent* retournent respectivement les nombres de tokens et de phrases. Ces traits permettent de reconnaître les SE présentant un contenu textuel étendu.

- **f_amorce** La seconde famille de traits est uniquement appliquée à l'amorce. Le trait *t_Saturation_s* indique si la dernière phrase de l'amorce est complète syntaxiquement, c'est-à-dire saturée (Bush, 2003). Une amorce incomplète indique généralement que « les constituants manquants sont fournis par un ou des items » (Maurel *et al.*, 2002)[5]. Le trait *t_Ponctuation* retourne la ponctuation terminant l'amorce. Il s'agit essentiellement de vérifier la présence d'une virgule, qui dénote une continuité de la phrase d'amorce. Les traits de type *t_Lexique* sont calculés en projetant des lexiques sur l'amorce. Ces lexiques correspondent notamment aux introducteurs et aux organisateurs de Bush (2003) (p. ex. *suivant*, *liste de*, etc.). Les circonstants sont également recherchés (p. ex. *En 1996*, etc.), ainsi que certains marqueurs de type de relation (p. ex. les verbes *être*, *composer*, *définir*, etc.).

Les SE (7.b) et (7.c) exemplifient l'application de traits. Les tokens entre crochets sont ceux capturés par *t_POS_p*. Les tokens soulignés sont ceux pris en compte par *t_POS_c* et *t_Lexique*. L'absence de phénomènes (p. ex. verbe dans l'item, etc.) est tout autant informative. Notons que la SE (7.b) porte une hyperonymie.

Traits	Informations capturées.
f_amorce_item	
t_POS_c	Booléen indiquant si une catégorie morpho-syntaxique donnée est présente dans l'amorce et dans l'item.
t_POS_p	Retourne les catégories morpho-syntaxique de début et de fin de l'amorce et de l'item.
t_NbToken	Nombre de tokens dans l'amorce et dans l'item.
t_NbSent	Nombre de phrases dans l'amorce et dans l'item.
f_amorce	
t_Saturation_s	Booléen indiquant si la dernière phrase de l'amorce est complète syntaxiquement (c'est-à-dire est saturée).
t_Ponctuation	Retourne la ponctuation terminale de l'amorce.
t_Lexique	Booléen indiquant si les tokens contenus dans un lexique donné sont présents dans l'amorce.

TABLE 7.1 : Synthèse des traits pour la qualification de la relation sémantique

[5] Cité par Porhiel (2007).

(7.b)

[Liste] des principaux estuaires de [France]

- [Estuaire] de la Garonne appelé aussi estuaire de la [Gironde] ;
- Estuaire de la Seine ;
- Estuaire de la Loire, partie aval de la Basse-Loire correspondant à l'embouchure de la Loire ;
- Estuaire de la Rance : voir aussi l'Usine marémotrice de la Rance ;
- La série des « Estuaire picards » à la configuration géomorphologique particulière (avec du sud au nord, les estuaires de la Somme, de la Canche, de l'Authie, de la Liane (artificialisé), de la Slack et du Wimmereux.

(7.c)

[L']Aquaculture en [France]

- [La] France a une tradition ancienne (plus de 1000 ans) de pisciculture extensive en étangs (Limousin, Dombes et nombreux viviers créé par les moines, et utilisation extensive des retenues de moulins dont les vers de farine et déchets de meunerie alimentaient les truites et d'autres poissons ainsi sédentarisés). Au début du XXe (Statistiques 2002, publiées en 2003) Environ 6 000 exploitants d'étangs déclarés, surtout localisés en Région Centre et Rhône-Alpes et Lorraine ont livré 12 000 tonnes (6 790 pour le repeuplement et 2 570 pour la consommation) de carpe, gardon, brochet et tanche, pour un chiffre d'affaires d'environ 16 millions d'euros. 80 % de la production part à la consommation directe, 12 % servent aux rempoissonnements pour la pêche de loisir et 8 % pour le repeuplement des [rivières].
- La salmoniculture en rivière puis la pisciculture marine sont plus récentes. 60 000 tonnes de poissons étaient produites par an au début des années 2000 (en 2002), pour environ 222 millions d'euros de chiffres d'affaires. salmoniculture (133,8 millions de chiffres d'affaires) a permis de produire environ 41 000 tonnes de truites arc-en-ciel (Bretagne et Aquitaine surtout). 52 producteurs en mer ont livré 5 800 tonnes, 3 000 tonnes de bar, 1 200 tonnes de dorade royale et 910 tonnes de turbot.
- La conchyliculture (huîtres, moules et coquillages) s'est fortement développée sur la façade atlantique.
- les conchyliculteurs ont produit 90 300 tonnes d'huîtres, 4 100 tonnes d'autres coquillages, produites par 52 600 concessions sur le domaine public sur 18 100 hectares et 1570 km de littoral.

7.2.2 Évaluation

Dans cette section, nous évaluons notre méthode pour la relation d'hyperonymie sur notre corpus annoté. Deux points sont discutés :

- Une évaluation quantitative,
- Une analyse des traits utilisés.

Évaluation quantitative Nous décrivons ici l'évaluation des trois tâches T_Onto, T_Hypo_1 et T_Hypo_2 sur le corpus annoté. Notre corpus de 745 SE a été scindé aléatoirement en deux parties. 80% des SE ont été utilisés comme ensemble de développement. Les 20 % restant ont été utilisés comme ensemble de test. Les trois tâches ont été évaluées face à une baseline consistant à classer les SE dans le type sémantique ciblé. Cette baseline est plus difficile à battre qu'une baseline aléatoire, car les distributions des classes positives et négatives sont asymétriques dans le corpus (Section 6.2.3).

Les informations morpho-syntaxiques ont été ajoutées avec l'analyseur syntaxique en dépendances Talismane[6] (Urieli, 2013). Pour les classifieurs, nous avons utilisé la librairie OpenNLP[7] pour la régression logistique et la librairie LIBSVM[8] pour le SVM avec noyau gaussien. Le perfectionnement des traits et le choix des hyper-paramètres ont été effectués par validation croisée (k=10) sur l'ensemble de développement.

La table 7.2 présente les résultats obtenus pour la tâche T_Onto, qui identifie les SE porteuses de relations de type *à visée ontologique*. La distribution des types sémantiques dans l'ensemble de test implique une bonne précision, et donc une bonne exactitude, pour la baseline. La régression logistique et le SVM montrent un gain. Toutefois, étant donné le nombre limité d'observations, seul le SVM montre un effet (p-valeur $< 0{,}03$ avec test t pairé), avec un Δ de 2,71 pour le F_1-score et un Δ de 6,17 pour l'exactitude.

Tâche	Configurations	Précision	Rappel	F_1-score	Exactitude
T_Onto	Régression logistique	80,65	93,46	86,58	78,77
	SVM	79,84	96,26	**87,29**	**79,45**
Baseline	Majorité	73,28	100,0	84,58	73,28

TABLE 7.2 : Résultats pour l'identification du type sémantique *à visée ontologique* dans la tâche T_Onto

La table 7.3 présente les résultats obtenus pour les tâches T_Hypo_1 et T_Hypo_2, qui identifient les SE porteuses d'une relation d'hyperonymie. Dans les deux cas, les résultats obtenus surpassent significativement la baseline majoritaire. L'ajout des sorties de la tâche T_Onto dans notre architecture en cascade montre un gain significatif pour T_Hypo_1 par rapport à T_Hypo_2. La table 7.4 résume les comparaisons faites avec un test t pairé.

[6] http://github.com/urieli/talismane
[7] http://opennlp.apache.org/
[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Tâches	Configurations	Précision	Rappel	F_1-score	Exactitude
T_Hypo_1	Régression logistique	78,01	84,78	81,25	**75,34**
	SVM	74,77	90,22	**81,77**	74,66
T_Hypo_2	Régression logistique	70,59	78,26	74,23	65,75
	SVM	71,05	88,04	78,64	69,86
Baseline	Majorité	63,01	100,0	77,31	63,01

TABLE 7.3 : Résultats pour l'identification de la relation sémantique d'hyperonymie dans les tâches T_Hypo_1 et T_Hypo_2

Comparaisons			p-valeurs
T_Hypo_1 Régression logistique	vs.	T_Hypo_2 Régression logistique	$< 0{,}01$
T_Hypo_1 Régression logistique	vs.	Baseline	$< 0{,}01$
T_Hypo_1 SVM	vs.	T_Hypo_2 SVM	$< 0{,}02$
T_Hypo_1 SVM	vs.	Baseline	$< 0{,}01$

TABLE 7.4 : Comparaisons entre les résultats obtenus pour l'identification de la relation sémantique d'hyperonymie

Les différences obtenues entre la régression logistique et le SVM ne sont pas significatives dans T_Hypo_1. Nous pouvons observer que la régression logistique atteint la meilleure précision, tandis que le SVM montre le meilleur rappel. Ces résultats étaient attendus, car la frontière de décision du SVM avec un noyau gaussien, plus flexible, semble être légèrement biaisée par les valeurs extrêmes. Ceci augmente le nombre de faux positifs dans les données de test.

Une observation des résultats a révélé que la régression logistique et le SVM semblaient apprendre des représentations différentes du problème. Une perspective immédiate d'amélioration de notre système consisterait à combiner ces classifieurs en parallèle dans des approches ensemblistes tels que le bagging ou le boosting (Zhou, 2012).

Analyse des traits Nous proposons ici une première analyse des traits afin de déterminer ceux qui semblent discriminants pour la reconnaissance de l'hyperonymie dans T_Hypo_1 et T_Hypo_2. Cette analyse a été faite sur l'ensemble de développement.

Le choix de représentation du problème comme un problème de classification binaire, nous permet d'utiliser une corrélation de Pearson. Pour un trait f donné, nous calculons celle-ci comme suit :

$$r_f = \frac{\text{cov}(\mathbf{f}, \mathbf{y})}{\sigma(\mathbf{f})\sigma(\mathbf{y})} \qquad \text{(eq.7.3)}$$

où $\mathbf{f}$ est un vecteur binaire où chaque dimension est l'application du trait f sur l'ensemble de développement. Le vecteur $\mathbf{y}$ est un vecteur binaire qui représente les classes positives et négatives de l'ensemble de développement. $\mathbf{f}$ et $\mathbf{y}$ sont de même longueur. La fonction cov(.,.) calcule la covariance des vecteurs donnés en argument. La fonction $\sigma(.)$ retourne l'écart-type du vecteur donné en argument.

Une valeur positive indiquera que le trait f est corrélé à la présence d'une relation d'hyperonymie portée par la SE, tandis qu'une valeur négative indiquera l'inverse. La table 7.5 ordonne les 10 traits présentant les plus grandes valeurs absolues de corrélation. La majorité d'entre eux sont liés à l'item.

Traits	Informations capturées	Composants	corrélation r
t_POS_c	contient : Verbe conjugué	Item	-0,259
t_POS_p	commence par : Déterminant	Item	0,235
t_NbToken	nombre tokens : 5	Item	0,147
t_POS_c	contient : Nom propre	Item	0,132
t_POS_p	commence par : Nom	Item	0,128
t_POS_c	contient : Nom pluriel	Amorce	0,120
t_POS_c	contient : Nom propre	Amorce	0,120
t_POS_p	commence par : Verbe infinitif	Item	-0,113
t_Lexique	marqueurs de relation : *metalexicale*	Amorce	-0,112
t_NbToken	nombre tokens : 3	Item	0,107

TABLE 7.5 : Ordonnancement des dix traits avec les valeurs absolues de corrélation les plus élevées pour la relation d'hyperonymie

La forme et les caractéristiques morpho-syntaxiques de l'item apparaissent être des sources d'information importantes pour déterminer si une SE porte une relation d'hyperonymie. Par exemple, les items qui contiennent un verbe conjugué ou qui débutent par un verbe infinitif sont négativement corrélés. *A contrario*, les items avec peu de contenu textuel sont positivement corrélées à la présence d'une relation d'hyperonymie.

De manière transversale, les informations relatives à la ponctuation s'avèrent être peu corrélées avec la relation d'hyperonymie. Seuls certains phénomènes spécifiques et en petit nombre apportent un gain d'information. Par exemple, la présence d'une virgule terminale dans l'amorce présente une corrélation de -0.05 avec l'hyperonymie. La SE (ul_{11}, ul_{12}, ul_{13}) l'exemplifie dans (7.a).

Concernant la tâche T_Hypo_1, les sorties de la tâche T_Onto présentent des taux de corrélation élevés avec la présence d'une hyperonymie, avec respectivement 0,361 pour la régression logistique et 0,321 pour le SVM.

Notons que si la corrélation de Pearson est un bon indicateur, ce score n'est néanmoins pas directement lié aux performances des classifieurs sur l'ensemble de test (ou sur de nouvelles données). Il serait intéressant d'étendre cette analyse des traits avec d'autres mesures telles que l'entropie[9]. Notons que nous donnons un tableau d'analyse identique en Annexe C.4 pour la tâche T_Onto.

[9] L'entropie serait intéressante, car elle est notamment utilisée pour l'apprentissage des arbres de décisions avec l'algorithme C4.5 (Quinlan, 1993).

7.3 Identification des arguments de la relation

Dans cette section, nous décrivons la méthode d'extraction des arguments de la relation. Ensuite, nous reportons l'évaluation sur le corpus annoté.

7.3.1 Description

Nous décrivons ici une méthode pour l'identification des arguments de la relation portée par les SE d'intérêt (Section 7.1). Nous considérons que les termes ainsi que les entités nommées sont acceptables comme arguments, et nous les désignons par le terme d'entité textuelle (Section 1.1.3). Deux objectifs sont poursuivis :

1. Il s'agit d'identifier l'entité textuelle dans l'amorce qui impose une « contrainte d'identité sémantique sur les items » (Schnedecker, 2002)[10]. Selon les approches et les phénomènes mis en avant, cette entité textuelle peut porter différentes appellations : enuméraThème (Ho-Dac *et al.*, 2010), classificateur (Bush, 2003), classifieur (Porhiel, 2007) ou encore énumérable (Tadros, 1985)[11]. Selon la relation sémantique portée par la SE, d'autres appellations peuvent également être utilisées telles que l'hyperonyme ou l'holonyme.

2. Il s'agit d'identifier les entités textuelles co-énumérées dans l'énumération qui saturent le classificateur (Porhiel, 2007) et « apparaissent dans une relation d'égalité vis-à-vis » de celui-ci (Ho-Dac *et al.*, 2010). Selon la relation sémantique portée, d'autres appellations peuvent être utilisées telles que l'hyponyme ou le méronyme.

Nous proposons de réaliser simultanément ces deux objectifs au travers d'une tâche de prédiction structurée. Nous considérons l'ensemble des entités textuelles impliquées dans la relation comme un chemin identifiable parmi d'autres au sein d'un graphe. Le choix de ce chemin (c'est-à-dire des entités textuelles qui le composent) s'effectue sur des critères typographiques, dimensionnels, lexicaux et syntaxiques. Nous tirons avantage du fait que les entités textuelles co-énumérées dans une SE apparaissent généralement au travers d'un parallélisme textuel.

Dans la pratique, ce problème est considéré comme une recherche du chemin de moindre coût dans un graphe liant les entités textuelles de la SE. Notre système est composé de trois modules : (i) un module qui transpose chaque SE et ses entités en un graphe, (ii) un algorithme de recherche du chemin de moindre coût dans ce graphe, et (iii) une méthode d'estimation du coût des arcs. Pour ce dernier module, deux méthodes sont proposées : la première repose sur une approche distributionnelle et la seconde repose sur de l'apprentissage supervisé. La figure 7.5 schématise le système dans son ensemble.

[10] Cité par Bras *et al.* (2008).
[11] Cité par Porhiel (2007).

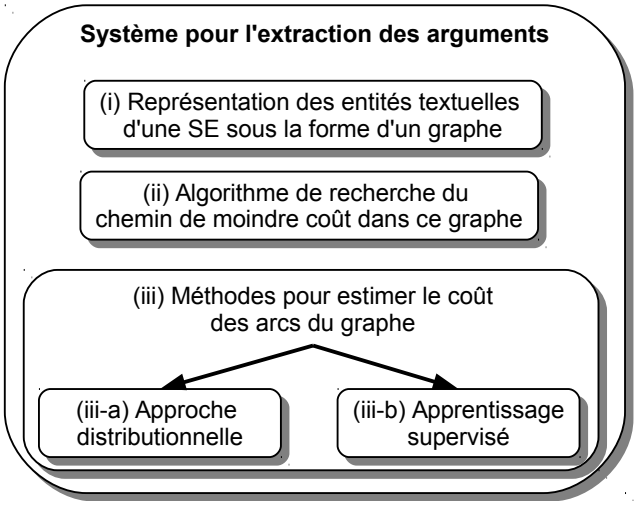

FIGURE 7.5 : Schéma du système d'extraction des arguments de la relation

Dans cette section, nous allons successivement décrire chacun des composants :
(i) le module de transposition des SE en graphes ;
(ii) l'algorithme de recherche de chemin ;
(iii) deux méthodes d'estimation du coût des arcs :
(iii-a) une fondée sur une approche distributionnelle ;
(iii-b) une fondée sur de l'apprentissage supervisé.

(i) Transposition des SE en graphes Nous transposons chaque SE d'intérêt en un graphe représentant les liens entre ses entités textuelles. Cela est fait en deux étapes :

1. Les entités textuelles de chaque SE d'intérêt sont identifiées : une analyse morpho-syntaxique est effectuée, et ensuite l'utilisation de patrons et d'analyseurs terminologiques permettent le balisage des entités textuelles. Un processus de post-traitement élimine les doublons et fusionne les entités textuelles qui se chevauchent.

2. Chaque SE d'intérêt de m items est représentée par un graphe acyclique dirigé, où les nœuds sont les entités textuelles et les arcs sont les liens possibles entre celles-ci. Ce graphe est décomposé en niveaux, où chaque niveau i contient les nœuds représentant les entités textuelles de l'unité logique ul_i ($0 \leq i \leq m$). Ainsi, le niveau 0 correspond aux entités textuelles de l'amorce, tandis que les niveaux supérieurs correspondent aux entités textuelles des items. Chaque nœud de niveau i ($0 \leq i < m$) est connecté par des arcs dirigés vers tous les nœuds du niveau $i+1$. Un nœud factice *racine* est ajouté au sommet du graphe.

À titre d'exemple, nous donnons un exemple de SE (7.d)[12], dont les entités textuelles sont soulignées, et son graphe correspondant en figure 7.6.

[12] Pour la clarté de l'exemple, deux items ont été enlevés. L'exemple complet est en Annexe D.1.

(7.d)

Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets...) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants. Les matériaux extraits sont transportés par une marie-salope.
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.

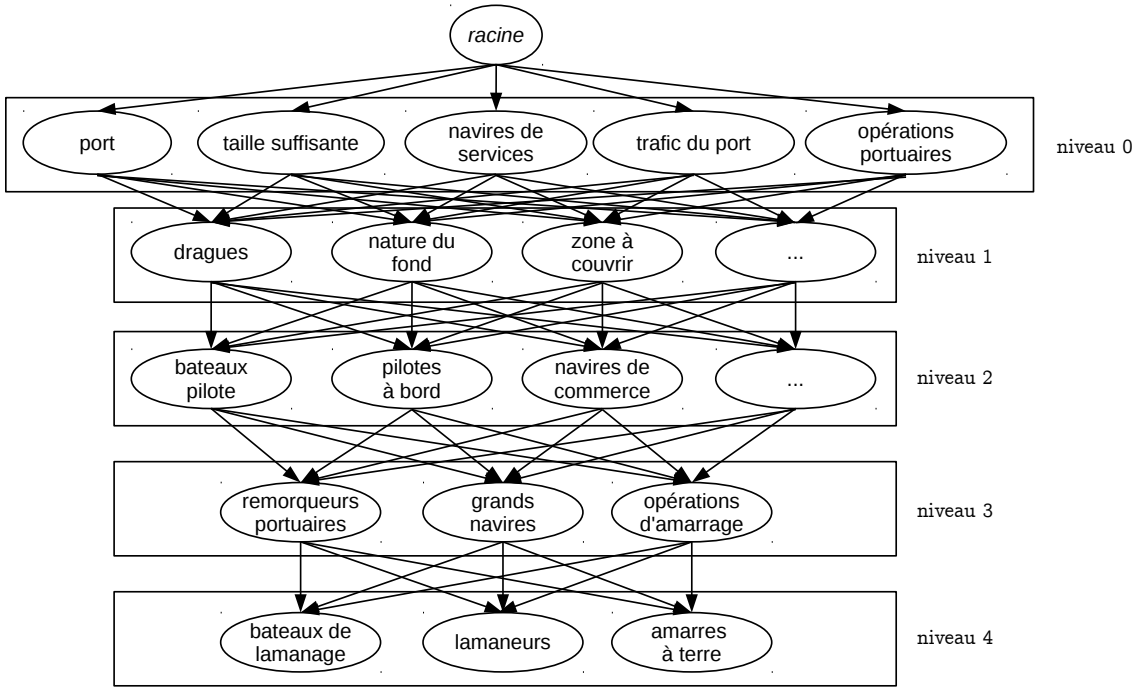

FIGURE 7.6 : Extrait de la représentation en graphe correspondant à l'exemple (7.d)

(ii) Algorithme de recherche du chemin de moindre coût La recherche des entités textuelles impliquées dans la relation s'effectue au travers d'une recherche du chemin de moindre coût entre la racine et les entités textuelles du dernier item. Le coût d'un arc est défini par :

$$\text{coût}(< T_i^j, T_{i+1}^k >) = 1 - \text{score}(T_i^j, T_{i+1}^k) \quad \text{(eq.7.4)}$$

où T_i^j est le j-ème nœud du niveau i. Nous cherchons à maximiser la fonction score(.,.), définie sur [0, 1], pour les paires d'entités textuelles impliquées dans la relation sémantique portée par la SE. Le calcul de ce score est discuté dans la section suivante.

À titre d'exemple, pour la SE (7.d) et son graphe (Figure 7.6), la séquence d'entités textuelles retournée devrait être ["navires de services", "dragues", "bateaux pilote", "remorqueurs portuaires", "bateaux de lamanage"]. Si nous identifions que cette SE est porteuse d'une relation d'hyperonymie, alors nous pouvons établir, par exemple, que "bateaux pilote" est un type de "navires de services".

La recherche du chemin de moindre coût est effectuée avec un algorithme A*, car celui-ci peut opérer sur de larges espaces de recherche avec l'heuristique adéquate. Dans l'algorithme 2, nous décrivons la procédure de recherche. La procédure prend en entrée le nœud *racine* et l'ensemble O des nœuds à atteindre. À chaque nœud est associé une séquence P d'arcs correspondant au chemin parcouru jusqu'à ce nœud dans le processus de recherche. L'estimation du coût des chemins dirige la recherche vers les nœuds les plus prometteurs. Le coût estimé pour un chemin P est défini par :

$$f(P) = g(P) + h(P) \quad \text{(eq.7.5)}$$

La fonction $g(P)$ calcule le coût réel du chemin P et est définie par :

$$g(P) = \sum_{<T_i^j, T_{i+1}^k> \in P} \text{coût}(< T_i^j, T_{i+1}^k >) \quad \text{(eq.7.6)}$$

L'heuristique $h(P)$ choisit, selon un principe glouton, le nouveau chemin au coût minimal sur d niveaux et retourne son coût :

$$h(P) = g(l_d(P)) \quad \text{(eq.7.7)}$$

La fonction $l_d(P)$ est définie récursivement. $l_0(P)$ est un chemin vide. Supposons que $l_d(P)$ est défini et $T_{i_d}^{j_d}$ est le dernier nœud atteint par le chemin formé par la concaténation de P et $l_d(P)$, alors nous définissons :

$$l_{d+1}(P) = l_d(P) \cdot < T_{i_d}^{j_d}, T_{i_d+1}^{p} > \quad \text{(eq.7.8)}$$

où p est l'index du nœud avec le cout d'arc minimal et appartenant au niveau $i_d + 1$:

$$p = \underset{k<|\text{niveau } i_d+1|}{\text{argmin}} \text{coût}(< T_{i_d}^{j_d}, T_{i_d+1}^{k} >) \quad \text{(eq.7.9)}$$

L'heuristique $h(P)$ est minorante, donc admissible.

Algorithme 2 Algorithme A* adapté pour la recherche du chemin de moindre coût dans un graphe acyclique. La procédure prend en entrée le nœud *racine* et l'ensemble O des nœuds du dernier niveau. T_i^j correspond au j-ème nœud du niveau i du graphe. À chaque nœud est associé une séquence P d'arcs correspondant au chemin parcouru jusqu'à ce nœud dans le processus de recherche. À chaque nœud correspond un ensemble de successeurs. β est la file de nœuds à trier selon l'estimation du coût des chemins associés (dite *liste ouverte*). Le graphe étant acyclique, une file des nœuds déjà parcourus (dite *liste fermée*) n'est pas nécessaire.

```
1: procedure ASTAR(racine, O)
2:     enfiler(racine, β)
3:     Tant Que β non vide Faire
4:         T_i^j ← défiler(trier(β))            //Tri selon le coût estimé des chemins
5:         Si estAtteint(O, T_i^j) :
6:             retourner(récupérerChemin(T_i^j))          //Retourne la solution
7:         Sinon
8:             Pour Chaque successeur de T_i^j Faire      //Continue la recherche
9:                 T_{i+1}^k ← copier(successeur)
10:                P ← récupérerChemin(T_i^j) · <T_i^j, T_{i+1}^k>   //Mise à jour du chemin
11:                associerChemin(P, T_{i+1}^k)
12:                enfiler(T_{i+1}^k, β)
13:            Fin Pour Chaque
14:        Fin Si
15:    Fin Tant Que
16: Fin procedure
```

Nous avons paramétré l'heuristique pour que celle-ci considère trois niveaux (d=3). Ce choix a montré qu'il s'agissait d'un bon compromis entre le nombre d'opérations et le nombre d'itérations durant la recherche. Si malgré cette heuristique la recherche n'aboutit pas après un nombre d'itérations arbitrairement grand, celle-ci est stoppée. Nous avons empiriquement établi ce nombre à 1000 itérations.

(iii) Méthodes pour l'estimation du coût entre deux entités textuelles

Nous proposons deux méthodes pour estimer le coût des arcs liant les entités textuelles du graphe (défini en eq.7.4). Ces méthodes cherchent à maximiser la fonction score(.,.) (i) entre le classificateur et la première entité textuelle énumérée, et (ii) entre les paires d'entités textuelles co-énumérées.

À titre d'exemple, pour la SE (7.d) et son graphe (Figure 7.6), le score obtenu pour les entités textuelles "dragues" et "bateaux pilote" devrait être plus élevé que celui obtenu pour la paire "dragues" et "navires de commerce".

Les deux méthodes utilisées pour l'estimation du score reposent respectivement sur une approche distributionnelle et sur de l'apprentissage supervisé.

(iii-a) Approche distributionnelle Cette méthode cherche à estimer le score en se fondant sur un critère de cohésion lexicale. Pour deux entités textuelles, la cohésion est mesurée au travers de la similarité cosinus entre leur vecteur respectif. Ceci nécessite préalablement la construction d'une représentation vectorielle des entités textuelles.

La construction de la représentation vectorielle est faite avec le modèle Skip-Gram proposé par Mikolov (2013b). Pour une séquence de mots $w_1, w_2, \ldots, w_T$, ce modèle cherche à prédire les mots qui entourent un mot w_t sur une fenêtre de taille $2c$. Dans ce contexte, la fonction objective prend la forme :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c}\log p(w_{t+j}|w_t) \qquad \text{(eq.7.10)}$$

Le calcul de $p(w_{t+j}|w_t)$ utilise un modèle exponentiel multinomial[13]. Après convergence de l'objective, les paramètres du modèle correspondent à la représentation vectorielle.

Nous estimons la cohésion lexicale entre deux entités textuelles données avec une similarité cosinus (eq.7.11). La fonction vec(.) retourne le vecteur correspondant à l'entité textuelle donnée en argument. Si l'entité présente plusieurs mots, la compositionnalité est traitée par simple addition des vecteurs des composants de l'entité textuelle. Enfin, nous bornons sur l'intervalle [0,1] la fonction de score (eq.7.12).

$$\text{sim\_cos}(T_i^j, T_{i+1}^k) = \frac{\text{vec}(T_i^j)\ \text{vec}(T_{i+1}^k)}{||\text{vec}(T_i^j)||\ ||\text{vec}(T_{i+1}^k)||} \qquad \text{(eq.7.11)}$$

$$\text{score}(T_i^j, T_{i+1}^k) = \begin{cases} 0 & \text{si sim\_cos}(T_i^j, T_{i+1}^k) \leq 0 \\ \text{sim\_cos}(T_i^j, T_{i+1}^k) & \text{sinon} \end{cases} \qquad \text{(eq.7.12)}$$

[13] Notons toutefois que le modèle hiérarchique introduit par Morin et Bengio (2005) est préféré à une régression logistique multinomiale pour des raisons de complexité. L'alphabet des mots pouvant être potentiellement large (10^5 à 10^9 mots), une normalisation n'est pas envisageable. Un problème similaire apparaît également avec les CRF (Annexe B).

(iii-b) Apprentissage supervisé Cette méthode cherche à estimer le score en se fondant sur des critères typographiques, dispositionnels, lexicaux et syntaxiques. Deux modèles sont utilisés : le premier intervient entre l'amorce et le premier item, tandis que le second intervient entre les paires d'items. Nous avons choisi d'utiliser des modèles probabilistes pour estimer le score directement par la probabilité obtenue. Dans ce cadre, la fonction de score prend la forme suivante :

$$\text{score}(T_i^j, T_{i+1}^k) = p(y|T_i^j, T_{i+1}^k) \quad \text{(eq.7.13)}$$

où y représente l'événement où la paire d'entités textuelles, représentées par T_i^j, T_{i+1}^k, est impliquée dans la relation sémantique portée par la SE d'intérêt.

Pour estimer la distribution de probabilité associée à chacun des deux modèles, nous utilisons une régression logistique. Cet algorithme d'apprentissage supervisé a déjà été présenté en section (7.2.1)[14]. Les traits exploités sont essentiellement les mêmes pour les deux modèles. Par contre, les phénomènes linguistiques appris étant distincts, les paramètres associés à ces traits sont différents dans chaque modèle. Le tableau 7.6 synthétise les traits. Quatre familles peuvent être distinguées :

- **f_contexte** Ces traits informent quant au contexte d'une entité textuelle donnée. Ceci permet de capturer des phénomènes tels que les organisateurs introduisant le classificateur, la présence de caractères de ponctuation, le parallélisme positionnel entre les entités textuelles co-énumérées, etc.
- **f_entité** Ces traits informent quant aux caractéristiques internes des entités textuelles. Les marques de pluriel ou la présence d'un nom propre sont prises en compte au travers des informations morpho-syntaxiques capturées. Un trait d'inclusion lexicale est également ajouté.
- **f_document** Ces traits exploitent la représentation en dépendances du document (Chapitres 4 et 5). Par exemple, il est vérifié si la SE d'intérêt est imbriquée ou imbrique d'autres structures textuelles. La position dans le document et les étiquettes logiques sont également prises en compte.
- **f_cosinus** Le trait proposé dans cette famille exploite la similarité cosinus permise grâce à la représentation vectorielle précédemment construite.

Pour éviter un déséquilibre entre les observations positives (paires d'entités textuelles impliquées dans la relation) et celles négatives, nous avons procédé à un sur-échantillonnage des observations positives pour obtenir un ratio 1:1. Cela a été fait selon une démarche aléatoire, comme proposé par Estabrooks *et al.* (2004).

[14] Notons qu'une présentation, plus complète, est également faite dans l'annexe B.

Traits	Informations capturées.
f_contexte	
t_POS_c	Contexte morpho-syntaxique de l'entité textuelle.
t_Position_c	Position de l'entité textuelle dans l'unité logique.
f_entité	
t_POS_e	Informations morpho-syntaxiques de l'entité textuelle.
t_Inclusion_e	Booléen indiquant la présence d'une inclusion lexicale.
t_NbCar_e	Nombre de caractères de l'entité textuelle.
t_NbToken_e	Nombre de tokens dans l'entité textuelle.
f_document	
t_Logique	Retourne l'étiquette logique de l'unité logique traitée.
t_Position_d	Position d'une unité logique dans l'ensemble du document.
t_Coord_Sub	Informations, pour une unité logique donnée, concernant la présence de coordonnés ou de subordonnés.
t_NbSent_d	Nombre de phrases dans une unité logique.
f_cosinus	
t_Sim_Cos	Similarité cosinus des vecteurs d'entités textuelles.

TABLE 7.6 : Traits pour l'identification des arguments des relations sémantiques portées par les structures énumératives d'intérêt

7.3.2 Évaluation

Nous évaluons notre méthode sur le corpus annoté. Deux points sont discutés :

- Une évaluation quantitative de la tâche ;
- Une analyse des traits utilisés.

Évaluation quantitative La méthode pour l'identification des entités textuelles (termes et entités nommées) impliquées dans des relations sémantiques de notre corpus est évaluée ici. Il s'agit en majorité de relations hiérarchiques (hyperonymie et holonymie) et, plus minoritairement, de relations de type *ontologique_autre*. Ce choix est fait car les entités textuelles impliquées dans ces relations montrent des caractéristiques communes d'apparition. Le corpus présentait 1511 paires d'entités textuelles. Les documents du corpus ont été scindés aléatoirement en deux parties de manière à obtenir deux tiers des paires d'entités textuelles comme ensemble de développement. Les documents contenant le tiers restant ont été utilisés comme ensemble de test.

Les informations morpho-syntaxiques ont été ajoutées avec Talismane (Urieli, 2013). Pour l'identification des entités textuelles, nous avons utilisé les extracteurs de termes ACABIT[15] (Daille, 1996) et YaTeA[16] (Aubin et Hamon, 2006). Pour chacun de ces outils, nous avons écrit un client Java permettant leur utilisation avec l'ensemble d'étiquettes

[15] http://www.bdaille.com/
[16] http://search.cpan.org/~thhamon/Lingua-YaTeA/

de Talismane[17] [18]. En post-traitement, nous avons enlevé les entités textuelles qui apparaissent en position isolée dans les amorces plus de deux fois dans notre corpus. Ceci permet de mettre de côté les titres génériques de Wikipédia[19] (p. ex. *liens externes*, etc.) et certains circonstants (p. ex. *France*, etc.). La liste complète (20 entités textuelles) est donnée en annexe C.5. Dans leur travail, Shinzato et Torisawa (2004a) (Section 1.3.2) proposent une solution similaire, mais construisent la liste à la main.

Nous avons utilisé l'implémentation word2vec[20] du modèle Skip-Gram (Mikolov *et al.*, 2013a) et le corpus FrWac[21] (Baroni *et al.*, 2009) de 1,6 milliard de mots pour l'approche distributionnelle. Deux représentations vectorielles ont été construites : la première de dimension 500, la seconde de dimension 200. Nous avons utilisé la librairie OpenNLP pour la régression logistique. Des études ont montré que les informations importantes étaient généralement disposées en début d'item (Ho-Dac, 2007). Nous proposons donc une baseline qui sélectionne les premières entités textuelles de chaque item. Le classificateur choisi est la dernière entité textuelle dans l'amorce.

Une mesure de confiance inverse est calculée pour chaque solution retournée par le système. Cette mesure correspond à la moyenne des coûts des arcs constituant le chemin retourné comme solution. Les résultats sont triés selon cette mesure et seuls ceux en deçà d'un seuil donné sont pris en compte. Pour chaque configuration, nous présentons les résultats correspondant au seuil maximisant le $F_{0,5}$-score. Ce choix est fait dans un contexte d'extraction de relations où nous voulons légèrement favoriser la précision sur le rappel. La table 7.7 présente ces résultats, avec additionnellement le F_1-score. La figure 7.7 montre l'ensemble des courbes précision-rappel pour tous les seuils.

Configurations	Précision	Rappel	$F_{0,5}$-score	F_1-score
Régression logistique	78,98	**69,09**	**76,78**	**73,71**
Similarité cosinus (dim, 500)	**83,71**	30,10	61,72	44,28
Similarité cosinus (dim, 200)	66,52	30,10	53,56	41,45
Baseline	48,37	69,09	51,46	56,91

TABLE 7.7 : Résultats pour l'identification des arguments de la relation sémantique

L'approche par régression logistique montre les meilleurs résultats. Les approches reposant sur la similarité cosinus montrent des scores de précision intéressants. Ceci semble confirmer un lien entre la cohésion lexicale des entités textuelles et leur implication dans une relation sémantique. Il apparaît également qu'augmenter la dimension des vecteurs améliore la précision, sans toutefois impacter le rappel. La baseline présente un rappel intéressant. Ceci confirme la tendance à trouver les informations saillantes en début d'item. Les erreurs de cette baseline concernent majoritairement les classificateurs.

[17] https://github.com/fauconnier/acabit-client
[18] https://github.com/fauconnier/yatea-client
[19] https://fr.wikipedia.org/wiki/Aide:Plans_d'articles
[20] https://code.google.com/p/word2vec/
[21] http://wacky.sslmit.unibo.it/doku.php?id=corpora

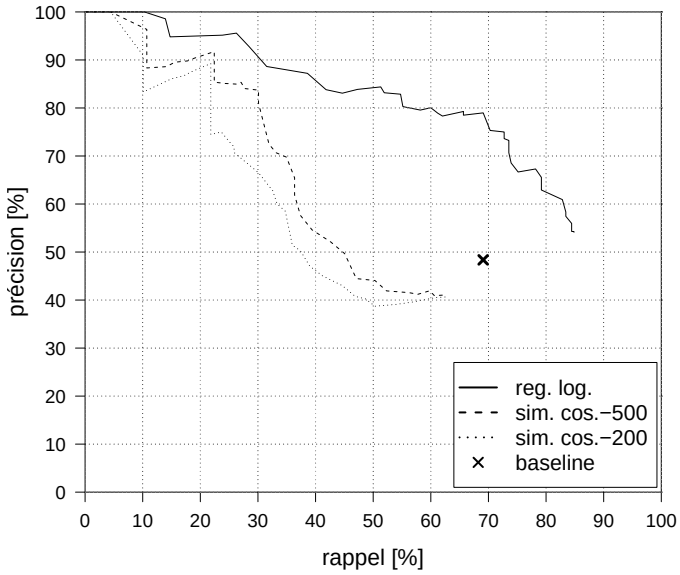

FIGURE 7.7 : Comparaison entre les configurations pour l'identification des arguments de la relation sémantique

Analyse des traits Nous proposons une analyse des traits pour l'approche par apprentissage supervisé. Ceci permet de mesurer l'implication des familles de traits dans l'identification des entités textuelles impliquées dans la relation. Cette analyse a été faite sur l'ensemble de test.

La table 7.8 présente les résultats lorsque nous enlevons chaque famille de traits. Nous voyons que la prise en compte du contexte et des informations internes des entités textuelles améliore les résultats. Nous constatons également que lorsque les traits relatifs à la structure de document sont enlevés, la précision est améliorée mais le rappel obtient une valeur inférieure à celle de la baseline. Ce résultat confirme l'intérêt de prendre en compte la structure de document pour l'identification des arguments des relations hiérarchiques dans les SE d'intérêt. *A contrario*, la suppression du trait de similarité cosinus ne fait pas baisser sensiblement les résultats. Ceci semble conforter la possibilité d'identifier les arguments en utilisant uniquement une approche endogène au texte.

Familles de traits	Précision	Rappel	$F_{0,5}$-score	F_1-score
-f_contexte	68,84	49,09	63,71	57,31
-f_entité	78,21	56,57	72,65	65,65
-f_document	**87,79**	53,74	77,91	66,67
-f_cosinus	78,67	67,07	76,04	72,41
Tous	78,98	**69,09**	**76,78**	**73,71**

TABLE 7.8 : Analyse des traits pour la qualification de la relation

7.4 Évaluation de l'ensemble du système

Dans cette section, nous proposons une première évaluation pour l'ensemble du système sur de nouvelles données. L'objectif visé est l'identification de la relation d'hyperonymie avec ses arguments. Les raisons avancées sont les mêmes que pour l'expérience menée pour la qualification de la relation sémantique (Section 7.2) : l'hyperonymie présente une fréquence d'apparition importante et généralement indépendante du domaine du corpus traité (Morin, 1999).

Les données utilisées pour cette évaluation proviennent de Wikipédia. Deux corpus ont été construits à partir des articles de deux domaines : *Transport* et *Informatique*. Pour chaque domaine, nous avons sélectionné aléatoirement 400 pages appartenant à la catégorie du domaine, ou à des catégories connexes. Pour chaque page, les traitements suivants ont été exécutés :

1. Analyse logique adaptée au langage de balisage WikiText (Chapitre 5) ;
2. Identification des SE d'intérêt (Section 7.1) ;
3. Qualification de la relation sémantique (Section 7.2) ;
4. Identification des arguments de la relation (Section 7.3).

Pour la qualification de la relation sémantique et l'extraction des arguments, nous avons utilisé une régression logistique, car celle-ci a montré une bonne précision dans l'évaluation individuelle de ces tâches. Les résultats retournés par le système ont été triés selon leur mesure de confiance inverse. Finalement, nous avons manuellement évalué les 500 premières paires d'entités textuelles. Les courbes en figure 7.8 indiquent la précision obtenue. Pour les deux domaines, environ 300 paires hyperonymes-hyponymes ont été retrouvées avec une précision d'environ 60% pour le seuil le plus élevé.

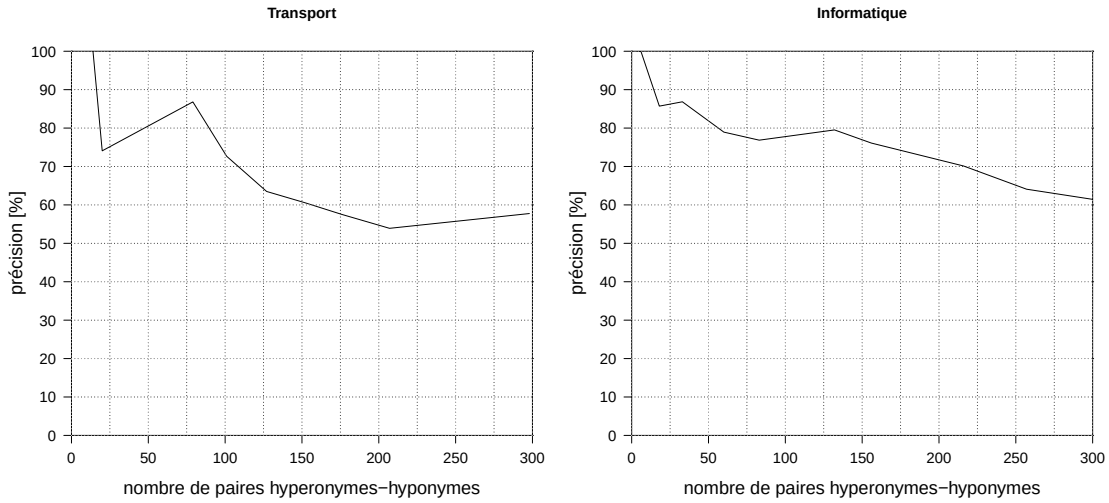

FIGURE 7.8 : Courbes de précision pour l'évaluation du système sur les domaines *Transport* et *Informatique*

Nous avons identifié plusieurs sources d'erreurs :

- **Imbrication des structures énumératives** La première source concerne les SE imbriquées. Dans ce cas, les unités logiques intermédiaires contiennent généralement des éléments contextuels sans présenter les hyponymes cherchés. Ces éléments contextuels peuvent être des caractéristiques additionnelles sur le classificateur. Bush (2003) parle de *differentia*. C'est notamment le cas dans l'exemple (7.e)[22]. Les entités textuelles soulignées sont celles retournées par le système.

  Les éléments contextuels peuvent également être des circonstants (temporels, spatiaux ou toute autre dimension) tel que dans l'exemple (7.f). Ce type de configuration est très présent dans les pages de *listes* de Wikipédia[23].

  Ces SE imbriquées constituent un aspect limitatif de notre solution. Une perspective consisterait à étendre la représentation faite pour l'extraction des arguments.

(7.e)

- transmission sans fil
  - Courte distance
    - Bluetooth
  - Moyenne distance
    - Wi-Fi, 802.11
    - MANET
  - Longue distance
    - MMDS
    - SMDS
    - Transmission de données sur téléphone cellulaire
    - Réseaux de téléavertissement

(7.f)

Pays de la Loire

- Ligne de Clisson à Cholet
- Ligne de Commequiers à Saint-Gilles-Croix-de-Vie
- Ligne de La Flèche à Vivy
- Ligne de La Possonnière à Niort
- Ligne de Nantes-État à La Roche-sur-Yon par Sainte-Pazanne et Commequiers (partiellement déclassée)
- Ligne de Saint-Hilaire-de-Chaléons à Paimbœuf
- Ligne de Saint-Nazaire au Croisic
- Ligne de Sainte-Pazanne à Pornic
- Ligne de Savenay à Landerneau par Redon et Quimper
- Ligne de Tours à Saint-Nazaire

[22] Certains items de troisième niveau ont été enlevés pour la clarté de l'exemple. L'exemple complet est accessible à l'annexe D.9.

[23] Voir la page Liste des listes : https://fr.wikipedia.org/wiki/Wikipédia:Liste_des_listes

- **Confusions avec d'autres relations sémantiques** La seconde source d'erreurs réside dans la difficulté à faire la distinction entre la relation d'hyperonymie et d'autres relations de type *à visée ontologique*. Ces erreurs interviennent essentiellement au niveau du module de qualification de la relation (Section 7.2). Deux groupes de confusions peuvent être distingués :
  - Le premier groupe concerne la relation d'holonymie. Cette relation est relativement fréquente dans les SE. La confusion est notamment faite car peu d'indices de surface signalent cette relation. La SE (7.g) l'exemplifie.
  - Le second groupe concerne les autres relations ontologiques. Généralement, ces relations sont marquées par un verbe qui permet de signaler leur nature. Par exemple, dans la SE (7.h) la relation peut être déterminée par le verbe « représente » terminant l'amorce. Notons au passage la difficulté pour le traitement des acronymes.

(7.g)

- Direction générale de l'aviation civile : supervise le bureau des enquêtes chargé des investigations sur les accidents et incidents aériens graves survenant sur le territoire national
  - Service des affaires générales
  - Bureau des enquêtes
  - Direction des études et de l'exploitation du transport aérien
  - Direction du personnel aéronautique et du matériel volant
  - Direction de la navigation aérienne

(7.h)

- Association des chemins de fer sud-africains (SARA, Southern African Railway Association), qui représente :
  - CFB (Chemin de fer de Benguela en Angola)
  - Botswana Railway
  - CFM (Chemins de fer du Mozambique)
  - Malawi Railway
  - Central East African Railway in Malawi
  - TransNamib
  - Swaziland Railway
  - Tazara (Tanzania/Zambia Railway Authority)
  - Zambia Railway
  - Tanzania Railways Corporation
  - NRZ (National Railways of Zimbabwe)
  - Beitbridge Bulawayo Railway
  - Metrorail d'Afrique du Sud
  - Spoornet (Afrique du Sud)

- **Phénomènes linguistiques complexes** Une autre source d'erreurs provient de phénomènes linguistiques. Ceux-ci montrent trop de variabilité pour qu'il soit envisageable de les traiter statistiquement avec le jeu de données à notre disposition. Citons notamment les phénomènes d'ellipse, d'anaphore, de négation, de coréférence, etc. La SE (7.i) montre un exemple où le classificateur n'est pas énoncé après l'organisateur. La SE (7.j) montre un exemple de négation.

(7.i)

Des volcanologues français divisent grosso modo les volcans du monde en deux types généraux :

- les « volcans rouges » aux éruptions effusives relativement calmes et émettant des laves fluides sous la forme de coulées. Ce sont les volcans de « point chaud », et les volcans d'« accrétion » principalement représentés par les volcans sous-marins des dorsales océaniques ;
- les « volcans gris » aux éruptions explosives et émettant des laves pâteuses et des cendres sous la forme de nuées ardentes ou coulées pyroclastiques et de panaches volcaniques. Ils sont principalement associés au phénomène de subduction comme les volcans de la « ceinture de feu du Pacifique ».

(7.j)

- membre supérieur : en dehors des traumatismes bénins, on retrouve :
  - fracture de la palette humérale
  - rupture de la coiffe des rotateurs
  - fracture de la clavicule
  - fractures du poignet, le plus souvent du scaphoïde

7.5 Discussion

Dans ce chapitre, nous avons présenté une méthode automatique pour extraire les relations sémantiques des SE verticales. Cette méthode est constituée de trois étapes.

Dans un premier temps, nous avons cherché à identifier parmi l'ensemble des SE celles qui étaient paradigmatiques et marquées dispositionnellement et typographiquement. Pour ces structures textuelles particulières, nous avons employé le terme de SE d'intérêt. Leur identification est effectuée par filtrage de motifs lors du parcours de l'arbre de dépendances représentant la structure des documents. Les résultats de cette méthode dépendent uniquement de la bonne construction préalable de cet arbre.

Dans un second temps, nous avons cherché à qualifier la nature des relations portées par les SE d'intérêt. Cette qualification a été effectuée sur la base de critères typographiques, lexicaux et syntaxiques. Pour la relation d'hyperonymie en particulier, un système d'apprentissage en cascade a été proposé. Cette architecture permet d'utiliser

les sorties d'un classifieur entraîné pour reconnaître les relations sémantiques de type *à visée ontologique* comme informations complémentaires pour le classifieur cherchant à identifier l'hyperonymie seule. Les résultats obtenus sont encourageants. L'analyse des traits a révélé que les traits lexicaux et syntaxiques intervenaient davantage que les traits typographiques pour déterminer la nature de la relation.

Dans un troisième temps, nous avons cherché à identifier les arguments impliqués dans les relations sémantiques portées par les SE d'intérêt. Cette identification a été effectuée sur la base de critères typographiques, dispositionnels, lexicaux et syntaxiques. Un système de prédiction structurée a été proposé. Celui-ci cherche à identifier le classificateur dans l'amorce et les entités textuelles qui sont en relation avec celui-ci. Ce problème est considéré comme une recherche de chemin de moindre coût. Une évaluation a été proposée pour les entités textuelles impliquées dans les relations hiérarchiques. Les résultats obtenus sont encourageants. L'analyse des traits a montré que les caractéristiques typographiques et dispositionnelles permettaient d'améliorer sensiblement les performances.

Une première évaluation de l'ensemble du système a été proposée pour la tâche d'extraction de la relation d'hyperonymie uniquement. Cette évaluation a été conduite sur deux domaines de Wikipédia. Les résultats ont montré que la précision était relativement correcte pour les premières paires hyperonyme-hyponymes retournées. Une analyse qualitative des erreurs a montré certaines sources d'erreurs. Celles-ci concernent notamment l'imbrication des SE, la confusion avec d'autres relations sémantiques (essentiellement l'holonymie) et les phénomènes linguistiques complexes (p. ex. ellipse, coréférence, etc.). Nous donnons, additionnellement, des exemples de SE dans l'annexe D.

Notre approche s'est penchée essentiellement sur la relation d'hyperonymie avec des méthodes statistiques. Cela est envisageable car cette relation est fréquente et habituellement exprimée indépendamment des domaines des corpus où elle apparaît. Ce point est sensible car il touche au problème de la régularité statistique des indices. Dans ce contexte, le traitement de relations sémantiques de domaine nécessiterait une adaptation de notre module de qualification de la nature de la relation. Généralement, pour ce type de cas, la nature de la relation est déterminée par le verbe porté par l'amorce, qu'il serait nécessaire d'extraire. Une approche par regroupement de prédicats verbaux, à la manière de Faure et Nédellec (1999), pourrait être envisagée. Par contre, le module d'extraction des arguments pourrait être utilisé sans qu'il soit nécessaire de lui apporter des modifications.

Les travaux les plus proches du nôtre sont ceux de Sumida et Torisawa (2008), précédemment introduits en section 1.3.2. Les auteurs considèrent un ensemble limité de balises (titres, listes à puces, listes ordonnées et listes de définitions) sur Wikipédia et proposent de lier le contenu textuel (appelés *title*) de ces balises de manière à obtenir des paires d'hyperonymes-hyponymes. Le traitement est fait à très grande échelle et seuls les cas simples et non ambigus sont retenus.

Notre travail diffère au moins en deux points. Premièrement, notre approche se veut générique en travaillant sur une représentation logique du document indépendante du format d'entrée. Dans ce contexte, l'intérêt est porté sur les phénomènes hiérarchiques, et non sur une sémantique préalablement associée à des balises. Deuxièmement, notre approche se veut plus fine en analysant le contenu textuel à la recherche des entités textuelles impliquées dans la relation sémantique. Ceci a nécessité la mise en place d'un système de prédiction structurée qui exploite la mise en forme du document. Sur ce dernier point, notre approche est particulièrement novatrice.

Conclusion et perspectives

Les systèmes d'extraction de relations sémantiques à partir de textes reposent généralement sur des méthodes qui n'exploitent pas tout le potentiel des textes : l'analyse se limite à un niveau phrastique et les éléments de mise en forme ne sont pas pris en considération. Or, nous avons vu qu'il était possible d'étendre ces méthodes afin d'extraire des relations exprimées au-delà des frontières de la phrase.

Dans ce cadre, nous nous sommes intéressés à la sémantique véhiculée par les indices typographiques et dispositionnels, et nous avons regardé dans quelle mesure ceux-ci pouvaient être exploités pour extraire des relations sémantiques. En particulier, notre étude s'est penchée sur les structures énumératives verticales, qui constituent un terrain idéal pour l'étude des interactions entre la mise en forme et le contenu sémantique.

Afin de permettre l'identification de ces structures textuelles dans les documents, nous avons montré qu'il était nécessaire de s'appuyer sur un modèle de représentation de la structure des documents offrant une abstraction de la mise en forme. Ensuite, pour cibler les structures énumératives verticales porteuses de relations sémantiques utiles à la construction de ressources, nous avons souligné la nécessité d'une réflexion linguistique sur des données attestées en corpus. Enfin, sur la base d'indices typographiques, dispositionnels, lexicaux et syntaxiques mis au jour en corpus, nous avons regardé dans quelle mesure il était possible d'extraire les relations au travers de deux étapes visant respectivement à qualifier la nature des relations portées par les structures énumératives et à identifier les arguments impliqués dans les relations.

Contributions

Au regard des objectifs énoncés, nous considérons quatre axes de contributions :

Modèle pour la structure hiérarchique des documents Nous avons proposé un modèle permettant la représentation de la structure hiérarchique des documents. Celui-ci offre une abstraction de la mise en forme, nécessaire face à la variabilité des indices visuels, ainsi qu'une connexion forte avec l'aspect rhétorique. Ce modèle se positionne dans la lignée des modèles théoriques rendant compte de l'architecture textuelle (Power *et al.*, 2003; Bateman *et al.*, 2001; Virbel, 1989), mais il s'en démarque en s'inscrivant dans une perspective d'analyse des textes.

Ceci est permis en articulant les unités logiques du document selon un principe de dépendance, et non un principe de composition comme cela est habituellement fait. Deux avantages peuvent être distingués : (i) il n'est plus nécessaire de définir des étiquettes abstraites avec des règles complexes d'inclusion, et (ii) la représentation offre une vue synthétique de l'organisation des documents, facilitant ainsi l'identification et la manipulation de structures textuelles hiérarchiques.

Un système d'identification de la structure logique Ce système implémente le modèle de représentation de la structure hiérarchique des documents sous la forme d'une méthode ascendante. Considérer l'analyse de la structure logique d'une manière équivalente à un problème d'analyse syntaxique a déjà été proposé dans la communauté d'Analyse du Document (Mao *et al.*, 2003). Toutefois, utiliser explicitement l'analyse en dépendances et sa représentation pour traiter la structure logique des documents est, à notre connaissance, nouveau. Ceci permet notamment d'utiliser des techniques éprouvées en analyse syntaxique (Nivre, 2008) et en parsing rhétorique (Hernandez et Grau, 2005). Une évaluation a été proposée sur un corpus de documents PDF. Les résultats obtenus sont encourageants et montrent qu'il est possible d'éliciter une représentation en dépendances des documents à partir des indices typographiques et dispositionnels qu'ils présentent.

Une typologie des structures énumératives, un outil d'annotation et un corpus annoté Pour caractériser et cibler les structures énumératives porteuses de relations sémantiques utiles à la construction de ressources, nous avons proposé une typologie considérant les dimensions visuelle, rhétorique, intentionnelle et sémantique. Cette typologie se veut complémentaire à celle de Luc (2000) et orthogonale à celle de Ho-Dac *et al.* (2010). Nous avons utilisé notre typologie pour établir un schéma d'annotation. Celui a été utilisé dans une campagne d'annotation utilisant l'outil d'annotation LARAt. Le corpus résultant a été exploité pour obtenir un retour quantitatif sur la typologie, mais également pour mettre au jour des indices de natures typographique, dispositionnelle, lexicale et syntaxique utiles dans la mise en œuvre du processus d'extraction de relations. La typologie, l'outil LARAt ainsi que les données annotées sont réutilisables par l'ensemble de la communauté dans le cadre des licences libres apposées.

Un système d'extraction de relations exploitant les structures énumératives paradigmatiques verticales Nous avons développé un système qui exploite le modèle de représentation de la structure du document et l'analyse linguistique des structures énumératives pour en extraire des relations sémantiques. Exploiter les éléments de mise en forme pour l'extraction de relations a déjà été proposé par d'autres approches, telle que celle de Sumida et Torisawa (2008). Néanmoins, notre approche présente deux avantages : (i) elle se veut plus générique en se basant sur une abstraction de la mise en forme, et non sur une sémantique préalablement associée aux balises, et (ii) elle se veut plus fine en analysant le contenu textuel à la recherche des entités textuelles impliquées dans la relation sémantique. Ceci a nécessité la mise en place de plusieurs modules d'apprentissage supervisé,

dont le dernier, qui vise à identifier les arguments de la relation, repose sur une technique de prédiction structurée. Sur ce dernier point, notre approche est particulièrement novatrice. Notre système a été évalué sur notre corpus annoté, ainsi que sur de nouvelles données. Les résultats sont encourageants.

Perspectives

Nous envisageons plusieurs perspectives immédiates à notre travail :

Apprentissage de classes d'équivalences visuelles Nous avons souligné dans le chapitre 4, la difficulté à définir un ensemble fini d'étiquettes pour représenter la structure logique des documents. Idéalement, nous pensons qu'il serait plus consistant de ne pas utiliser d'étiquettes, mais plutôt de proposer des classes d'équivalences visuelles. Ceci tiendrait sous l'hypothèse que, pour un document et un auteur (ou groupe de co-auteurs) donnés, une même mise en forme amène un même rôle logique. Des premières expériences ont été menées dans cette direction à l'occasion de l'encadrement d'un stage étudiant (terminé en octobre 2015). Les résultats préliminaires sur des articles de TALN Archives (Boudin, 2013) avec un algorithme de clustering incrémental (CobWeb) (Fisher, 1987) ont montré la faisabilité de l'approche. Toutefois, des mesures complémentaires sont encore nécessaires. Par exemple, il faudrait évaluer cette approche sur des documents courts.

Analyse visuelle des documents Web Dans son état actuel, le système proposé pour l'identification de la structure logique des documents prend en entrée, alternativement, des documents au format PDF ou au format WikiText. Nous avons effectué quelques expériences pour étendre le support au format HTML. Néanmoins, il est apparu que les documents Web étaient généralement très bruités : dans de nombreux cas, les rédacteurs Web privilégient la mise en forme visuelle au détriment du respect de la syntaxe HTML. Dans ce contexte, il paraît naturel d'envisager une analyse visuelle. Cette perspective nous rapprocherait des travaux de Manabe et Tajima (2015). Une première expérience pourrait consister à générer artificiellement des données d'entraînement en faisant varier les feuilles CSS d'un corpus relativement propre (p. ex. Wikipédia) et, ainsi, apprendre de grands modèles statistiques. Ce type de supervision distante nécessiterait une parallélisation sur une plate-forme de calcul distribué.

Structures énumératives imbriquées Dans le cas des structures énumératives imbriquées, généralement les unités logiques intermédiaires contiennent des éléments contextuels (p. ex. circonstants) et ne présentent pas complètement les arguments dans la relation. Dès lors, ceux-ci sont hors d'atteinte de notre système. Afin de dépasser cette limite, il serait nécessaire de modifier le graphe orienté acyclique utilisé pour l'identification des arguments. Une solution pourrait consister à agréger dans un même niveau i du graphe toutes les entités textuelles présentes dans l'unité logique i avec les entités textuelles des unités logiques subordonnées à i.

Cela nécessiterait la mise en place d'heuristiques simples (p. ex. ne prendre que les n premières entités textuelles d'une unité logique) afin de limiter la taille de l'espace de recherche.

Extraction de relations de domaine Dans cette thèse, nous avons principalement travaillé sur la relation d'hyperonymie. Celle-ci est relativement fréquente dans les structures énumératives (Gala, 2003). Dès lors, il est possible d'emprunter une approche statistique pour apprendre les régularités qu'elle présente. Néanmoins, des résultats antérieurs ont montré que ce type d'approche était difficile à mettre en œuvre pour des relations avec peu d'occurrences (Fauconnier *et al.*, 2013b). Une perspective à notre travail consisterait à traiter les relations de domaine (Grabar *et al.*, 2004), plus rares, mais qui sont généralement marquées lexico-syntaxiquement dans l'amorce des structures énumératives. Une approche par regroupement de prédicats verbaux, à la manière de Faure et Nédellec (1999), pourrait être envisagée.

Modélisation conceptuelle des relations Dans notre travail, notre attention s'est portée sur l'expression linguistique des relations sémantiques et sur leur repérage automatique dans les textes à partir d'indices de surface. Dans ce cadre, la phase d'interprétation au cours de laquelle le lien est fait entre la modélisation linguistique et la modélisation conceptuelle (Hirst, 2009) n'a pas été faite. Ce passage des entités textuelles aux concepts pourrait constituer une perspective à notre travail. Cela nécessiterait de (i) transposer les relations extraites dans un format exploitable par une machine (p. ex. XML/RDF), et de (ii) les organiser sous la forme de hiérarchies (taxonomies) ou de graphes (ontologies). Cette perspective nous rapprocherait des travaux récents proposés à SemEval pour la tâche TExEval (*Taxonomy Extraction Evaluation*) (Bordea *et al.*, 2015).

Ouverture à d'autres objets textuels La méthode proposée dans ce travail de thèse pourrait être appliquée à d'autres objets textuels, c'est-à-dire d'autres segments textuels rendus perceptibles à la surface des textes par leur mise en forme (Luc, 2000). En substance, il s'agit de mettre en évidence, après observations en corpus, des indices typographiques et dispositionnels caractéristiques. L'abstraction de ceux-ci au travers d'un modèle de représentation de la structure logique des documents facilite la mise en place de procédures d'identification. Ensuite, une fois un objet textuel identifié dans la structure du document, un travail d'analyse peut être débuté pour extraire des connaissances de celui-ci. Par exemple, il serait intéressant de voir dans quelle mesure les définitions, qui montrent des caractéristiques semblables à celles des structures énumératives (Péry-Woodley, 2000), pourraient être exploitables dans un contexte d'extraction de relations.

Annexes

Annexe A

Planches de documents

Sommaire

Ces planches présentent des observations faites dans les corpus LING et GEOP. En particulier, nous illustrons les différents types d'unités logiques élémentaires qu'il est possible d'y trouver et les difficultés associées à leur étiquetage logique à cause de la variabilité des indices visuels.

A.1 Extrait de ling_corbin

Cet extrait du document lin_corbin contient un titre de document, trois bylines, un titre de niveau 1 (h1) et un paragraphe. Le dernier bloc byline, une épigraphe qui cite un auteur, exemplifie la difficulté à faire une distinction nette entre mise en forme visuelle, rôle logique et rôle discursif.

Quel avenir pour la lexicographie française ?

Pierre Corbin

Université Charles de Gaulle - Lille 3
UMR 8163 “Savoirs, Textes, Langage”
pierre.corbin@univ-lille3.fr

L’investissement financier que suppose ce genre de produits est relativement important ; il dépasse de loin les moyens du chercheur isolé. Il exige, soit une décision proprement politique, soit la recherche capitaliste d’une rentabilité.
Rey (2008 : 13)

1 Introduction : d’une utopie humaniste au rêve de Sue Atkins

Les dictionnaires sont des textes importants. Témoignant de ce qui s’est déjà dit ou écrit pour guider ce qui pourra se dire ou s’écrire, ils reflètent, par leurs contenus, leur diffusion et leurs usages, les rapports d’une culture à son idiome ou les relations qu’elle entretient avec d’autres cultures et l’intérêt qu’elle porte à leurs idiomes[1]. Mais ces liens sont tout sauf simples, et leurs reflets sont volontiers brouillés. S’ils sont peu diversifiés et pauvres en substance, les dictionnaires constituent pour leurs destinataires des repères faciles et d’utilisation aisée mais laissant sans réponses nombre de questions ; s’ils sont plus variés et plus riches, donc plus complexes, leur choix adéquat requiert du discernement et leur utilisation, moins immédiate, demande application et patience[2]. La difficulté à trouver une information dans un dictionnaire, surtout dans une version imprimée de celui-ci, étant susceptible de croître avec la probabilité qu’elle y figure, ces répertoires deviennent d’autant plus élitistes que leur matière s’enrichit et que le traitement de celle-ci s’affine : attestant simultanément de la vitalité des idiomes dont ils traitent et de l’attachement que vouent à ceux-ci certains locuteurs, mais se désancrant *ipso facto* du rôle utilitaire qui est au principe de cette classe d’ouvrages, ils tendent alors à trouver leur fin dans leur propre développement, ce qui les prédispose à être salués comme des œuvres dont le nom s’inscrira dans la liste des monuments de la lexicographie à côté d’autres produits de l’esprit sélectionnés pour l’admiration et l’exégèse[3], en même temps que se restreint le nombre de ceux qui, étant disposés à assumer le coût de leur acquisition et les efforts requis par leur consultation, peuvent assez maîtriser celle-ci pour en tirer profit.

FIGURE A.1 : Extrait du document ling_corbin

A.2 Extrait de geop_2

Cet extrait du document geop_2 montre le type de mise en forme visuelle présente dans le corpus GEOP. Dans ce document, l'interligne est très marqué et les titres ne sont pas numérotés. Le premier élément complique la segmentation en blocs textuels, tandis que le second complique l'analyse logique.

Introduction

Ce nouveau rapport du Centre français sur les Etats-Unis fait suite au *policy paper* sur « Le Contrôle de l'imagerie satellitaire, un dilemme américain », publié en septembre 2001, puis mis à jour dans une version en anglais en mars 2002.

L'objectif initial du présent rapport était de présenter les résultats d'une étude menée par le *National Security Council* (NSC) sur le contrôle de la diffusion de l'imagerie commerciale par le gouvernement américain. Lancée au printemps 2001, cette étude devait s'achever en début d'année 2002.

Malheureusement, certaines difficultés structurelles et conjoncturelles sont apparues et cette étude officielle n'a pas encore vu le jour. Le groupe de réflexion chargé de l'étude était une sous-commission particulière du *Policy Coordinating Committee* sur l'espace (*PCC-space*) créé par le NSC. Dans les faits, les efforts du NSC en matière spatiale n'ont pas été suffisants. L'autorité au sein des sous-groupes n'était pas clairement attribuée au NSC. Ed Bolton, *Director for Space* au NSC sous l'autorité de Frank Miller, n'était pas de rang suffisant pour imposer des compromis aux différentes agences réunies dans le *PCC-Space*. Surtout, les événements du 11 septembre ont axé les priorités du gouvernement sur l'action et non sur la réflexion.

FIGURE A.2 : Extrait du document geop_2

A.3 Extrait de ling_roche

Cet extrait du document ling_roche montre des blocs d'en-têtes et un pied de page.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Morphologie
DOI 10.1051/cmlf08064

suffixe *-eur* sur un verbe ou un nom d'activité. Application systématique de ce principe dans la série des dérivés en *-isme* et en *-iste* : l'existence, dans le même paradigme dérivationnel, d'un adjectif de relation en *-ien* bloque la formation d'un dérivé en *-iste* (la forme en *-ien* en tient lieu) et réciproquement l'existence d'un dérivés en *-iste* bloque généralement la formation d'un adjectif de relation en *-ien* (la forme en *-iste* en tient lieu), alors que les deux suffixes ne sont pas équivalents (cf. Roché, 2007).

(8a) *Staline* → *stalinien* Adj 'de Staline'
↗ " N Adj 'partisan de Staline', 'favorable à Staline''
Staline → °*staliniste*

(8b) *Lénine* → *léniniste* N Adj 'partisan de Lénine', 'favorable à Lénine'
↗ " Adj 'de Lénine'
Lénine → °*léninien*

Ici encore, la dimension lexicale – l'influence du paradigme dérivationnel – l'emporte sur la logique constructionnelle.

2.2 Un paradigme lexical unifié

Revenons aux ethniques pour observer maintenant le paradigme lexical qu'ils constituent. Lorsqu'ils sont formés sur le nom de pays, ils sont marqués comme tels soit par un suffixe spécifique – *-ais* ou *-ois* le plus souvent (*Congolais, Béninois*), *-(i)ote* (*Chypriote*), *-ite* (*Yéménite*), *-i* (*Emirati*), etc. –, soit par un suffixe dont c'est un emploi privilégié – *-ien* et *-ain* principalement (*Canadien, Cubain*). Lorsque le nom de peuple est le primitif, en revanche, sa forme est imprévisible (*Russe, Turc, Arabe…*) et ne permet pas de le caractériser. D'où la tendance à doter le nom de peuple originel d'une finale suffixale qui va l'intégrer dans le paradigme : *Anglais* se substitue à *Angle, Danois* à *Dane, Finnois* à *Finn, Hongrois* à *Hongre,* etc. Il s'agit là du phénomène d'« hypercaractérisation diachronique » depuis longtemps mis en évidence (cf. Malkiel, 1957-1958) : un lexème déjà caractérisé par son sens tend à l'être également par sa forme, si les deux se sont pas conjointement marqués d'emblée. Ou, en d'autres termes, d'une « intégration paradigmatique » (Corbin, 1987). Cette forme très particulière de suffixation – sémantiquement tautologique – est exactement semblable à l'exemple classique des noms d'arbres du type *peuplier* (formé sur l'afr. *peuple,* qui avait déjà le même sens). Dans certains cas, le suffixe sert directement d'habillage à un emprunt (*Malais, Iroquois, Illinois*), comme pour le non moins classique *palétuvier.* D'autres dénominations (*Gascon, Breton*) sélectionnent un ancien cas régime de préférence au cas sujet (*Gasc, Bret* sont également attestés) parce qu'il leur donne une apparence suffixale et les agrège à un paradigme en *-on* (*Wallon, Frison, Saxon, Teuton*) qui s'étend lui aussi par intégration paradigmatique (*Lapon, Letton*).

Une variante consiste à construire par suffixation un nouvel ethnique sur le nom de pays construit lui-même sur le primitif : *Malais* → *Malaisie* → *Malaisien*. D'où les nombreux doublets : *Finnois* / *Finlandais, Somali* / *Somalien, Azéri* / *Azerbaïdjanais, Thaï* / *Thaïlandais,* etc. Ils peuvent servir à distinguer le groupe ethnique proprement dit, d'un côté, et les citoyens d'un Etat, de l'autre : tous les Azéris ne sont pas Azerbaïdjanais et tous les Malaisiens ne sont pas Malais. Mais dans la pratique ils tendent à être interchangeables, avant que l'un chasse l'autre. On trouve sur la Toile de nombreuses attestations de « gouvernement thaï », « gouvernement malais », « gouvernement azéri » alors qu'on attendrait normalement « thaïlandais », « malaisien », « azerbaïdjanais ». Ainsi se continue sous nos yeux le processus historique qui a fait passer les ethniques du type (1) au type (2) puis au type (3). Un peuple sans territoire attitré (type (1)) s'installe dans un « pays » auquel il faut donner un nom (type (2)) ; et quand ce pays devient un Etat au sens moderne du terme les individus ne sont plus perçus comme les membres d'un peuple mais comme les habitants de cet Etat (type (3)). Si les *Espagnols* sont aujourd'hui les habitants de l'*Espagne,* l'Espagne (*Hispania*) a d'abord été le pays des *Hispani*. Et ainsi de l'Italie, de la France, etc.

La place manque pour étudier le rôle des différents suffixes dans ce double mouvement. D'une façon générale, leur distribution dépend en grande partie de critères phonologiques. Les bases en /i/ donnent à 88 % des dérivés en *-ien* (*italien, estonien, bolivien, tanzanien, fidgien…*). Que ces dénominations soient, souvent, empruntées à d'autres langues européennes n'y change rien : il est vraisemblable que les

FIGURE A.3 : Extrait du document ling_roche

A.4 Extrait de geop_24

Cet extrait du document ling_roche montre des notes de bas de page.

rise le présent, l'immédiat et le visible[8]. La démocratie comme culture n'est en revanche pas en prise avec le temps mondial, car elle a besoin de temps. Elle n'est de surcroît pas immédiatement ou clairement identifiable. Elle est toujours relative et, par-là même, contestable. La prise en charge de la question démocratique par la communauté internationale renforce encore ce hiatus. On peut à peu près dire si des élections dans tel ou tel pays sont libres – et, le cas échéant, stigmatiser les contrevenants ; en revanche, on peut difficilement évaluer la réalité d'une culture démocratique. On imagine fort bien une délégation du Congrès américain tancer tel ou tel dirigeant pour ne pas respecter la démocratie ; on l'imagine moins évaluant la culture démocratique de ce même pays.

Qu'est-ce que la mondialisation ?

Commençons tout d'abord par donner une définition succincte de la mondialisation qui constitue la toile de fond de notre propos. Il s'agit pour l'essentiel d'un processus d'intensification des relations sociales planétaires, qui se traduit par une disjonction croissante entre l'espace et le temps. Qu'est-ce à dire ? Que les lieux où se déroulent les événements sont géographiquement de plus en plus éloignés des lieux où leurs conséquences s'expriment. Dans une société traditionnelle, l'espace dans lequel vit et se meut l'individu est un espace physique généralement limité. Tout ce qu'il voit et tout ce qu'il fait a pour cadre le village où il est né. Son horizon spatio-temporel est donc très limité. Dans une société moderne, cet espace de référence s'élargit pour toute une série de raisons, dont la plus importante est la spécialisation croissante des rôles et des fonctions. À partir du moment où l'on ne fait pas tout soi-même, on est obligé de s'adresser à d'autres pour obtenir certains biens et services[9].

La mondialisation intervient donc comme un processus d'élargissement de l'espace de référence dans lequel les acteurs sociaux s'insè-

8. Voir les travaux décisifs de Juan Linz sur la relation entre temps et démocratie et, notamment, J. Linz et A. Stepan, *Problems of Democratic Transition and Consolidation*, Baltimore, Johns Hopkins University Press,1996. Voir aussi Philippe Schmitter et Javier Santiso, « Three temporal dimensions to the consolidation of democracy », *International Political Science Review*, 1998, vol. 19, n° 1, et Zaki Laïdi, *Le Temps mondial*, Bruxelles, Complexe, 1997.

9. Anthony Giddens, *Les Conséquences de la modernité*, Paris, L'Harmattan, 1994.

FIGURE A.4 : Extrait du document geop_24

A.5 Extrait de ling_deMulder

Cet extrait du document ling_deMulder montre des références bibliographiques. L'une d'entre elles présente ce qui peut s'apparenter à une erreur de mise en forme.

Références

Blank, Andreas (1997). Prinzipien des lexikalischen Bedeutungswandel am Beispiel der romanischen Sprachen. Tübingen: Niemeyer.

Bybee, Joan L., Perkins, R.D., et Pagliuca, W. (1994). The Evolution of Grammar : Tense, Aspect and Modality in the Languages of the World. Chicago : Chicago University Press.

Bybee, Joan (2006). « From usage to grammar : The mind's response to repetition ». Language 82/4, 711-733.

Detges, Ulrich (1999). « Wie entsteht Grammatik ? Kognitive und pragmatische Determinanten der Grammatikalisierung von Tempus Markern. » In Lang, Jürgen, et Neumann-Holzschuh, Ingrid, éds. Reanalyse und Grammatikalisierung in den romanischen Sprachen. Tübingen : Max Niemeyer Verlag, 31-52.

Flydal, Leiv (1943). Aller et venir de suivis de l'infinitif comme expressions de rapports temporels. Oslo: I Kommisjon Hos Jacob Dybwad.

Fries, Charles C. (1927). « The expression of the future. » Language 3, 87-95.

Gougenheim, Georges (1929 : 1971). Etude sur les périphrases verbales de la langue française. Paris : Nizet.

Hopper, Paul, et Traugott, Elizabeth (2003). Grammaticalization. Cambridge : Cambridge University Press. Deuxième édition.

Koch, Peter (1996). « Der Beitrag der Prototypetheorie zur Historischen Semantik: Eine kritiche Bestandsaufnahme ». Romanistisches Jahrbuch 46, 27-46.

Koch, Peter (1999). « Frame and contiguity. On the cognitive bases of metonymy and certain types of word formation ». In Panther, Klaus-Uwe, et Radden, Günter, éds. Metonymy in Thought and Language. Amsterdam / Philadelphia : John Benjamins, 139-167.

Koch, Peter (2004). « Metonymy between pragmatics, reference, and diachrony ». Metaphorik.de 07/2004.

Kuteva, Tania (2001). Auxiliation. An enquiry into the nature of grammaticalization. Oxford : Oxford University Press.

Leeman-Bouix, Danielle (1994). Grammaire du verbe français. Des formes au sens. Paris : Nathan.

Littré, Emile (1961/62). Dictionnaire de la langue française. Paris: Gallimard.

Nicolle, Steve (1998). « A relevance theory perspective on grammaticalization ». Cognitive Linguistics 9-1, 1-35.

Ruiz de Mendoza Ibáñez, Francisco José, et Hernández, Lorena Pérez (2003). « Cognitive operations and pragmatic implication ». In Panther, Klaus-Uwe, et Thornburg, Linda L., éds. Metonymy and Pragmatic Inferencing. Amsterdam / Philadelphia : John Benjamins, 23-49.

Sperber, Dan et Wilson, Deirdre (1995). Relevance : Communication and Cognition. Oxford : Basil Blackwell. Deuxième édition.

Traugott, Elizabeth C. & Dasher, Richard B. (2002). Regularity in Semantic Change. Cambridge: Cambridge University Press.

Werner, Edeltraut (1980). Die Verbalperiphrase im Mittelfranzösischen. Eine semantisch-syntaktische Analyse. Frankfurt a.M. : Lang.

Wilmet, Marc (1970). Le système de l'indicatif en moyen français. Genève : Droz.

Wilson, Deirdre (2006). « Pertinence et pragmatique lexicale ». Nouveaux cahiers de linguistique française 27, 33-52.

FIGURE A.5 : Extrait du document ling_deMulder

A.6 Extrait de ling_dal

Cet extrait du document ling_dal montre la distinction faite dans le corpus LING entre les références bibliographiques et les notes de fin de document. Ces dernières ne sont pas introduites préalablement par un titre.

Poitevin, P. (1879). *Nouveau dictionnaire de la langue française [...]*. Paris : C. Reinwald, libraire-éditeur.

Sander, E. (2000). *L'analogie, du naïf au créatif. Analogie et catégorisation*. Paris : l'Harmattan.

Saussure (de), F. (1916). *Cours de linguistique générale*. Paris : Payot, 1981.

Scalise, S. (1984). *Generative Morphology*. Dordrecht (Holland)/Cinnaminson (U.S.A.) : Foris Publications.

Schaar (van der), M. (1999). L'analogie et la vérité chez Franz Brentano. *Philosophiques, 26/2*, 203–217.

Singh, R. & Starosta, S. eds (2003). *Explorations in Seamless Morphology*. New Delhi/Thousand Oaks/London : Sage publications.

Skousen, R. (1989). *Analogical modeling of language*. Dordrecht : Kluwer Academic.

Skousen, R. (1992). *Analogy and structure*. Dordrecht : Kluwer Academic.

Skousen, R., Lonsdale, D. & Parkinson, D. B. eds. (2002). *Analogical Modeling. An exemplar-based approach to language*. Provo, Utah : Brigham Young University.

Stroppa, N. & Yvon, F. (2005). Apprentissage par analogie et rapports de proportion : contributions méthodologiques et expérimentales. *CAP 2005*, 61-62. [http://www.computing.dcu.ie/~nstroppa/papers/2005-CAP.pdf]

Touratier, C. (1988). Le problème des "lois phonétiques". Cercle linguistique d'Aix-en-Provence, *Travaux, 6*, 133-161.

Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960), 16 vol., Paris, Éditions du CNRS (t. 1-10) / Gallimard (depuis le t. 11), 1971-1994.

Vallès, T. (2004). *La creativitat lèxica en un model basat en l'ús*. Barcelona : Publicacions de l'Abadia de Montserrat.

---

[1] Le nom *analogie* est emprunté au latin *analogia* « rapport, conformité », lui-même emprunté au grec ἀναλογία « proportion mathématique », puis « correspondance, analogie » (*Le Grand Bailly*, 2000). Quintilien (dans *De Institutione Oratorie 1.6.*), pour qui l'un des fondements du langage que constitue la raison s'appuie principalement sur l'analogie, a proposé de traduire le terme par *proportio*.

[2] Aristote, *Métaphysique,* 1016b31, cité d'après Schaar (1999). On trouve une définition similaire dans *Poétique*, 21, d'après Milner (1989 : 631, n. 2).

[3] La définition du concept en mathématiques revient à Euclide (-3^e^ siècle), le fondateur de la géométrie, au sein d'une théorie de la proportion reprise à Eudoxe (-408, -355). Une application connue de ce principe en mathématiques est la « règle de trois », ou « règle de proportionnalité », définie comme la recherche, dans une proportion, du quatrième nombre, les trois autres étant connus.

[4] Biela (1991) mentionne encore l'utilisation qui est faite du concept dans les sciences naturelles, en anthropologie, en cybernétique, en ethnologie, dans le domaine juridique, en philologie, en sociologie, etc. A soi seul, l'article **analogie** du *Trésor de la langue française* donne un bon aperçu de la variété des domaines de spécialité qui recourent à la notion.

[5] Schématiquement, les analogistes appartenaient à l'école d'Alexandrie, tandis que les anomalistes étaient des stoïciens. Le débat passionna également les grammairiens latins : Varron lui consacra plusieurs livres de son *De Lingua latina* (*cf.* notamment le livre IX, consultable sous : http://www.udl.es/usuaris/s2430206/varroll2.htm), et on retrouve le thème chez Quintilien et César, pour ne citer qu'eux.

[6] *Cf.* aussi Furetière (1690, *s.v.* **analogie**) : « En Grammaire l'usage est souvent contraire à l'*analogie* des mots ».

FIGURE A.6 : Extrait du document ling_dal

A.7 Extrait de ling_gerard

Cet extrait du document ling_gerard montre une double imbrication de structures énumératives. La première indentation est marquée par un retrait visuel. La seconde indentation est marquée par la numérotation. Le dernier paragraphe joue le rôle de clôture.

Appuyée à ce dispositif théorique, la différenciation d'un style par les normes d'un genre se complique néanmoins de divers cas de figure, qu'on doit à l'existence de fonctionnements génériques multiples, dont certains mettent à l'épreuve l'analyse en composantes. C'est notamment le cas du poème en prose. D'ordinaire définit selon des critères très génériques comme un « texte poétique court, autonome et autotélique »[5], il apparaît dénué de prescriptions d'ordre thématique, narratif ou énonciatif, et ne définit donc pas d'interaction sociolectale au plan du contenu. Qu'une telle indétermination soit possible donne à imaginer, à titre heuristique, différentes façons pour un style de s'affirmer sous le régime textuel des composantes *sémantiques*. Dans cette perspective, les rapports entre genre et style rencontrent quatre cas de figure :

1. Interaction sociolectale définie

1a — On n'observe pas de régularités qui accuseraient une spécification idiolectale de l'interaction en question au sein du corpus d'étude (oeuvre littéraire, philosophique, etc.). Le style ne module pas son identité sur les prescriptions sémantiques du genre d'accueil et par conséquent la description des particularités individuelles devrait logiquement se poursuivre dans d'autres directions (cf. *infra* 3).

1b — On observe des régularités thématiques, narratives ou énonciatives justiciables d'un modèle individuel de production des textes. La description rend alors compte d'un style en tant qu'il se différencie sur un fond de caractéristiques génériques.

2. Interaction sociolectale indéterminée

2a — On n'observe pas d'interaction régulière entre composantes. À nouveau, il faudrait enquêter selon d'autres perspectives pour dégager les particularités individuelles recherchées.

2b — Le corpus présente en tout ou partie des régularités idiolectales significatives. Elles indiquent une forme d'appropriation individuelle où, pour une oeuvre donnée, la sémantique du genre le cède à la sémantique d'un style.

Les cas 2a et 2b correspondent à l'indétermination sémantique que nous avons illustrée avec le poème en prose. À l'inverse, les cas 1a et 1b présupposent des prescriptions au plan du contenu, *vis-à-vis* desquelles s'apprécient la singularité d'un mode de production et d'un mode d'anticipation du sens textuel (pour une réception familiarisée avec les textes ainsi mis en série). À cet égard, alors que 1b localise la différence de degré entre style et genre que signalent les propositions de Rastier, 2b réaliserait lui un investissement stylistique « catégoriel ». Les analyses suivantes illustrent ce dernier cas de figure sur un corpus poétique réduit.

FIGURE A.7 : Extrait du document ling_gerard

A.8 Extrait de geop_22

Cet extrait du document geop_22 illustre un cas de paragraphe qui débute sur une page et termine sur la suivante. Pour obtenir une représentation plane du document, ce type d'ambiguïté doit être résolu manuellement ou semi-automatiquement.

Les chances et la signification
d'une politique européenne de sécurité et de défense
dans le nouveau contexte international

Jean Klein

La crise provoquée par l'intervention armée contre l'Irak pour le contraindre à respecter les termes de la résolution 1441 du Conseil de sécurité avant que la mission des inspecteurs de l'United Nations Monitoring, Verification and Inspection Commission (UNMOVIC) soit achevée a ébranlé les fondations du système international et mis en évidence les « fractures » de l'Europe. En l'occurrence, des analystes n'ont pas hésité à voir dans la stratégie mise en œuvre par les Etats-Unis une violation des normes inscrites dans la charte de San Francisco et, dans leur penchant pour l'unilatéralisme, une contestation radicale de la responsabilité qui incombe à l'Organisation des Nations unies (ONU) pour le maintien et le rétablissement de la paix. L'Organisation du traité de l'Atlantique Nord (OTAN), dont la fonction initiale était la défense collective contre la menace soviétique et qui est demeurée, après l'effondrement de l'ordre bipolaire, l'une des principales organisations de sécurité dans « l'espace euro-atlantique » serait, elle aussi, vouée au dépérissement, les Américains préférant créer des alliances *ad hoc* pour défendre leurs intérêts et lutter contre le terrorisme (*coalition of the willings*) plutôt que de voir leur liberté d'action entravée par les contraintes d'une décision collective. Enfin, la construction d'une Europe de la défense serait compromise non seulement en raison de l'opposition de l'Administration de George W. Bush à la réalisation de ce projet, mais également du fait des divisions des Européens et de l'allégeance atlantique de la plupart des pays d'Europe centrale et orientale qui seront admis dans l'Union européenne (UE) en mai 2004[1].

Il ne saurait être question de vérifier le bien-fondé de ces jugements, ni de nous livrer à des spéculations sur la légalité de la guerre contre l'Irak ou de mesurer son impact sur la configuration du système international et les équilibres au Moyen-Orient. Notre propos est plus modeste et se bornera à l'examen des conséquences de cette crise sur l'organisation de la sécurité en Europe et l'avenir des relations transatlantiques. Il convient en effet de se demander si les divergences entre Européens qui se sont manifestées à cette occasion[2] annoncent une mutation radicale de la politique européenne de sécurité et de défense (PESD), sinon son abandon pur et simple, ou s'il ne s'agit que d'une crise passagère qui ne met pas en question les options fondamentales prises par l'UE après la guerre du Kosovo et entérinées par le Conseil d'Helsinki en décembre 1999. On sait que ce projet a suscité d'emblée des réserves de la part des Etats-Unis et que le président Bush ne lui a pas ménagé ses critiques lors de son premier voyage en Europe, en juin 2001 ; dans ses interventions au siège de l'OTAN et au Conseil européen de Göteborg (Suède), il reprocha notamment aux Européens l'insuffisance de leur effort de défense et dénonça leur prétention à mener une politique indépendante alors qu'ils n'en avaient pas les moyens. Le fait est que la plupart des Etats européens n'étaient pas prêts à faire les sacrifices nécessaires pour réduire l'écart entre leurs capacités militaires et celles des Etats-Unis et que la PESD restait un objectif lointain.

Quelques mois plus tard, les attentats terroristes de New York et de Washington créaient une situation nouvelle et donnaient lieu à l'expression d'une solidarité sans faille des Européens avec les Etats-Unis. L'article 5 du traité de l'Atlantique Nord fut invoqué à cette occasion et les actes terroristes perpétrés le 11 septembre 2001 sur le territoire américain furent qualifiés « d'attaque armée ». Mais, par un curieux paradoxe, la lutte contre les réseaux Al-Qaida et le régime des Talibans qui leur offrait un refuge en Afghanistan a été menée en dehors du cadre de l'Alliance et les Etats-Unis ont tenu pour quantité négligeable le concours de leurs alliés européens, à l'exception de celui du Royaume-Uni, qui a été associé dès l'origine à l'opération militaire baptisée « Liberté immuable » (*Enduring Freeedom*). D'aucuns ont interprété cette attitude comme la confirmation de la tendance à l'unilatéralisme américain et y ont vu le signe avant-coureur du dépérissement de la fonction militaire de l'alliance. D'autres, au contraire, ont souligné l'utilité de l'OTAN comme cadre de concertation des politiques de sécurité des Etats membres et

Communication présentée à l'Université de Keimyung (Corée du Sud), en mai 2003, dans le cadre de la conférence internationale « Past, Present and Future of European Integration: Issues and Prospects », organisée par The East Asian Association of International Studies, The Korean Society of Contemporary European Studies et l'Université de Keimyung.

[2] Le 30 janvier 2003 paraissait dans le *Wall Street Journal* une lettre ouverte dans laquelle cinq Etats membres de l'UE et trois pays candidats prenaient fait et cause pour la politique américaine vis-à-vis de l'Irak et prônaient la constitution d'un front uni entre l'Europe et les Etats-Unis. Le 5 février suivant, dix pays d'Europe centrale et orientale (le groupe de Vilnius) publiaient une déclaration qui allait dans

FIGURE A.8 : Extrait du document geop_22

A.9 Extrait de geop_31

Cet extrait du document geop_31 illustre un cas où deux titres de niveau 3 (h3) présentent une mise en forme comparable à des items (retrait visuel et puce). Ce type de confusion visuelle rend difficile l'étiquetage logique.

Le système GATT ne répond plus : limites du mercantilisme, évolution des rapports de force, nouveaux acteurs

- *Doit-on « payer » les règles de droit ? La méthode mercantiliste à l'épreuve.*

Depuis l'instauration du GATT, le libre-échange progressait aux rythmes de cycles de négociations, paradoxalement mus par le mercantilisme de l'échange de concessions

(...)

marchés publics et facilitation des échanges – et préférant poursuivre hors de l'OMC, par accords bilatéraux, les deux autres objectifs de régulation. Tous les pays développés avaient, par contre, un point commun : celui de refuser de « payer » par davantage de libéralisation agricole (impliquant des ajustements à coût politique immédiat élevé) l'élaboration de règles de droits (dont le bénéfice économique potentiel se diffuse à moyen ou long terme).

La méthode mercantiliste, issue des négociations du GATT, a rencontré à Cancun ses limites, pour traiter simultanément des enjeux de libéralisation et de régulation.

- *Certains deviendraient-ils aussi égaux que d'autres ? Le « consensus censitaire » à l'épreuve*

Lors de la création de l'OMC, les négociateurs pouvaient se référer à deux modèles de gouvernance. Celui de l'ONU, fondé globalement sur le « suffrage universel » et l'égalité des Etats à l'assemblée générale – sous réserve du Conseil de Sécurité – était aussi celui de l'ancien GATT. Celui des institutions économiques et financières de Bretton Woods était par contre fondé sur le « suffrage censitaire », lié au stock de capital détenu. Issus du GATT, qui était resté essentiellement un « club de riches » aux intérêts économiques comparables, la plupart de ces négociateurs admirait l'efficacité du deuxième système.

FIGURE A.9 : Extrait du document geop_31

A.10 Extrait de ling_abdoulhamid

Cet extrait du document ling_abdoulhamid illustre un cas de citation présentant une possible confusion avec un paragraphe. L'indentation logique n'est pas marquée par une indentation visuelle. Les indices utiles ici sont typographiques : (i) les deux points qui terminent le paragraphe qui précède la citation, et (ii) la présence de guillemets doubles autour de la citation.

1 Introduction

Les travaux descriptifs qui sont faits sur les propositions subordonnées circonstancielles concernent essentiellement celles qui ont un verbe conjugué à un mode personnel. Ils font abstraction des subordonnées participiales, (désormais SP) qui sont reconnues dans l'exemple suivant : *le chat parti, les souris dansent*. Beaucoup de grammaires de référence ignorent cette construction (Wagner et Pinchon 1991), d'autres la méconnaissent (Wilmet, 1997). Celles qui l'évoquent l'expliquent en la mettant en équivalence avec une subordonnée circonstancielle conjonctive en *dès que* ou *lorsque* : *dès que le chat est parti, les souris dansent* (Grevisse 1993 ; Riegel et al 1994).

Dans leur analyse, les grammaires qui parlent de la construction avancent l'idée que son procès est antérieur à celui de la proposition qui l'héberge (désormais PH), en précisant que la construction peut être indifféremment précédée par des éléments comme *une fois, sitôt, aussitôt,* et que le participe peut être précédé de l'auxiliaire *étant*. C'est également la position d'A. Borillo (2006 : 5) dans son étude sur les structures participiales à prédication seconde. Elle avance l'explication suivante :

« On peut constater que l'absence du marqueur temporel est parfois possible, sans réelle modification du sens de l'énoncé, si ce n'est que *sitôt*, *aussitôt,* et *à peine* ajoutent effectivement une précision d'immédiateté et que *une fois* souligne de manière explicite la relation d'antériorité d'une première éventualité par rapport à une autre. *Une fois le texte rédigé, il fallut le taper sur un stencil* ; *le texte rédigé, il fallut le taper sur un stencil.* Sans marqueur temporel, le sens reste très proche, de même que les règles de construction : le participe passé est celui d'un verbe construit avec le verbe *être*, qui doit être interprété avec une valeur passive si le verbe est transitif, avec une valeur active si le verbe est inaccusative ».

L'objectif de mon propos est de montrer que, pour mieux comprendre le fonctionnement de la SP, celle-ci doit être analysée, non pas dans le cadre de la phrase, mais dans le cadre du discours. En effet, comme B. Combettes (1993) l'a montré, les constructions détachées sont des éléments qui assurent la continuité thématique du discours. Elles reprennent, en général, des référents contenus dans le contexte antérieur. En tant que telle, la SP peut difficilement avoir un référent nouveau. Elle a en général un référent qui est déjà présent dans le discours. Il paraît donc difficile de se contenter de l'analyse phrastique pour rendre compte de ce type de construction. B. Combettes (1993 : 39-40) l'a souligné : « La construction détachée apparaît […] comme un constituant dont le fonctionnement dépend autant, sinon plus, de contraintes textuelles, de facteurs discursifs, que de caractéristiques strictement syntaxiques : le prédicat réduit qu'elle constitue se comporte en fait comme un prédicat intermédiaire, passage entre deux énoncés, qui prolonge le contexte de gauche dans une fonction de maintien d'un référent thématique. »

FIGURE A.10 : Extrait du document ling_abdoulhamid

Annexe B

Apprentissage supervisé

Sommaire

Cette annexe propose de présenter les algorithmes d'apprentissage supervisé utilisés dans ce travail. Une attention particulière a été donnée à la notation employée. Celle-ci se veut au possible consistante au travers de la présentation de tous les algorithmes.

Dans un premier temps, nous introduisons quelques notions préliminaires. Ensuite, nous présentons les algorithmes d'apprentissage supervisé. Une comparaison des algorithmes clôture cette annexe.

B.1 Notions préliminaires

B.1.1 Définitions générales

L'**apprentissage automatique** est une discipline regroupant un ensemble de techniques permettant à une machine d'apprendre, par elle-même, une fonction. Une fois apprise, cette fonction permet de prendre des décisions sur de nouvelles données. Cette discipline

est utilisée dans des situations pour lesquelles (i) le problème ne peut être résolu de manière analytique, mais où (ii) des données relatives à ce problème sont disponibles pour construire une solution statistique (Abu-Mostafa *et al.*, 2012).

Dans le cas de l'**apprentissage supervisé**, les exemples donnés pour l'apprentissage de la fonction sont *étiquetés*, c'est-à-dire qu'ils sont associés à une variable, continue ou discrète, dont la valeur indique la sortie attendue. Dans notre travail, nous appliquons l'apprentissage supervisé sur données étiquetées par une variable discrète. Il s'agit donc de *classification* : les étiquettes sont considérées comme des classes et l'objectif est de classer correctement les nouveaux exemples. Kotsiantis *et al.* (2007) expliquent :

> In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

L'apprentissage supervisé est défini par contraste à l'**apprentissage non supervisé** où les données ne sont pas préalablement étiquetées. Dans ce cadre, les algorithmes utilisés cherchent généralement à faire correspondre une sortie identique à des exemples partageant des propriétés similaires (Forgy, 1965).

Notons également l'**apprentissage par renforcement**. Dans ce paradigme, les algorithmes ne prédisent pas une étiquette, mais cherchent, dans un environnement contrôlé, un comportement qui est récompensé ou pénalisé par une score.

B.1.2 Composants de l'apprentissage supervisé

Formellement, l'objectif de l'apprentissage supervisé est d'approximer une **fonction cible** f qui transforme un espace d'entrée $\mathcal{X}$ en un espace de sortie $\mathcal{Y}$:

$$f : \mathcal{X} \rightarrow \mathcal{Y} \tag{eq.B.1}$$

L'espace $\mathcal{X}$ est l'ensemble de tous les vecteurs $\mathbf{x}$ et l'espace $\mathcal{Y}$ est l'ensemble de toutes les valeurs y. Le vecteur $\mathbf{x} \in \mathcal{R}^d$ est le vecteur de traits de chaque exemple et y est l'étiquette associée à chaque exemple. Les traits ou *prédicteurs*[1] représente une information de nature continue ou discrète, idéalement informative, pour un exemple donné.

Généralement, il est complexe de déterminer la fonction cible f analytiquement. L'apprentissage supervisé permet de l'approximer à partir des exemples et leur étiquette. Ces paires $(\mathbf{x}, y)$ constituent l'**ensemble d'apprentissage** noté $\mathcal{D}$ et de taille N :

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\} \tag{eq.B.2}$$

[1] « In the statistical literature the inputs are often called the predictors, a term we will use interchangeably with inputs, and more classically the independent variables. In the pattern recognition literature the term features is preferred, (...) » (Hastie *et al.*, 2009)

où y_n est l'application de la fonction cible $f(\mathbf{x}_n)$ aux exemples $n = 1, ..., N$. À partir de cet ensemble d'apprentissage, un **algorithme d'apprentissage** permet de choisir une formule $g : \mathcal{X} \to \mathcal{Y}$ qui approxime f, c'est-à-dire que nous avons $g \approx f$. Cette fonction g, appelée **hypothèse finale**, est choisie parmi un **ensemble d'hypothèses**, noté $\mathcal{H}$. La difficulté réside dans la nécessité que la fonction g puisse généraliser le comportement appris sur des exemples hors du corpus d'apprentissage. La figure B.1, adaptée de Abu-Mostafa *et al.* (2012), représente les liens entre les différents composants de l'apprentissage supervisé.

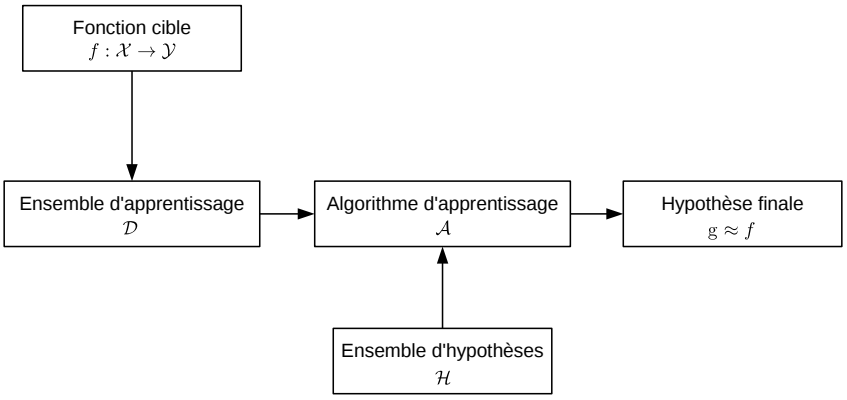

FIGURE B.1 : Relations entre les composants d'un problème d'apprentissage

Pour un problème donné, la fonction cible f et l'ensemble d'apprentissage $\mathcal{D}$ sont dépendants. Inversement, l'algorithme d'apprentissage et son ensemble d'hypothèses sont indépendants du problème. Dans la littérature, une certaine confusion apparaît entre les termes *modèle* et *algorithme d'apprentissage*. Dans ce travail, nous utilisons le terme d'algorithme d'apprentissage pour désigner $\mathcal{A}$ et nous utilisons les termes de modèle ou de classifieur pour désigner l'hypothèse obtenue au terme de l'apprentissage.

B.1.3 Composants de l'algorithme d'apprentissage

Généralement, un algorithme d'apprentissage $\mathcal{A}$ est décomposable en deux fonctions : (i) une fonction d'activation, et (ii) une fonction objective.

- La **fonction d'activation** d'un algorithme d'apprentissage correspond à la forme générale des hypothèses h dans l'ensemble $\mathcal{H}$. Cette fonction prend en argument un vecteur $\mathbf{x}$ correspondant aux traits d'un exemple et un vecteur de paramètres θ qui définit les poids associés aux traits. Par exemple, dans un algorithme d'appren-

tissage de type régression linéaire[2], l'hypothèse aura la forme :

$$h_\theta(\mathbf{x}) = \theta^{(1)}x^{(1)} + \theta^{(2)}x^{(2)} + \ldots + \theta^{(d)}x^{(d)} + b \qquad \text{(eq.B.3)}$$

où b reflète le terme biais. La fonction d'activation de la régression linéaire est une fonction affine. Pour simplifier la notation, le terme b sera ensuite représenté par le paramètre $\theta^{(0)}$. Ceci nécessite d'introduire dans $\mathbf{x}$ une variable $x^{(0)}$ valant 1 pour tous les exemples et, par conséquent, la dimension d est ajustée. Dans ce contexte, nous exprimons la fonction affine de la régression linéaire sous la forme d'une somme :

$$h_\theta(\mathbf{x}) = \theta^{(0)}x^{(0)} + \theta^{(1)}x^{(1)} + \theta^{(2)}x^{(2)} + \ldots + \theta^{(d)}x^{(d)} = \sum_{i=1}^{d} \theta^{(i)}x^{(i)} \qquad \text{(eq.B.4)}$$

- La **fonction objective**, aussi appelée fonction de coût (*cost function*) (Sammut et Webb, 2010), correspond à l'application d'un modèle h_θ sur l'ensemble d'apprentissage afin d'évaluer dans quelle mesure ce modèle approxime le comportement attendu sur les exemples observés :

$$\mathcal{L}(\theta) = diff(h_\theta(X), \mathbf{y}) \qquad \text{(eq.B.5)}$$

Au plus ce coût est élevé, au plus les sorties produites par le modèle diffèrent des valeurs observées dans l'ensemble d'apprentissage. Inversement, au plus ce coût est bas, au plus le modèle a appris les données observées.

B.1.4 Notation utilisée

Dans la suite de cette annexe, nous reprenons les conventions de notation relatives à l'algèbre linéaire. Nous représentons les traits d'un exemple donné sous la forme d'un vecteur colonne $\mathbf{x}$ de dimension $d + 1$ (avec le biais compris). Nous représentons les paramètres d'un modèle sous la forme d'un vecteur colonne θ de dimension $d + 1$. Ces deux vecteurs sont représentés ci-dessous :

$$\underbrace{\mathbf{x} = \begin{bmatrix} x^{(0)} \\ x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{bmatrix}}_{\text{Traits d'un exemple}}, \qquad \underbrace{\theta = \begin{bmatrix} \theta^{(0)} \\ \theta^{(1)} \\ \theta^{(2)} \\ \vdots \\ \theta^{(d)} \end{bmatrix}}_{\text{Paramètres du modèle}} \qquad \text{(eq.B.6)}$$

[2] La régression linéaire est un algorithme d'apprentissage supervisé qui prédit des valeurs continues. Nous l'utilisons ici comme exemple, car c'est un algorithme d'apprentissage qui présente une fonction d'activation simple illustrant le propos.

L'ensemble X des exemples d'apprentissage est représenté par une matrice de dimension $N \times (d+1)$ dont les lignes sont les vecteurs $\mathbf{x}_n$ transposés. Le vecteur $\mathbf{y}$ est un vecteur colonne représentant les valeurs y_n associées à chaque $\mathbf{x}_n$.

$$\underbrace{\mathrm{X} = \begin{bmatrix} \text{——}\mathbf{x}_1^{\mathrm{T}}\text{——} \\ \text{——}\mathbf{x}_2^{\mathrm{T}}\text{——} \\ \text{——}\mathbf{x}_3^{\mathrm{T}}\text{——} \\ \vdots \\ \text{——}\mathbf{x}_N^{\mathrm{T}}\text{——} \end{bmatrix}}_{\text{Ensemble d'entrée}}, \qquad \underbrace{\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix}}_{\text{Vecteur de sortie}} \tag{eq.B.7}$$

Avec cette notation, nous pouvons réécrire une fonction affine (eq.B.4) en l'exprimant sous la forme d'un produit scalaire des vecteurs de paramètres et de traits :

$$h_\theta(\mathbf{x}) = \sum_{i=1}^{d} \theta^{(i)} x^{(i)} = \theta^{\mathrm{T}}\mathbf{x} \tag{eq.B.8}$$

B.2 Algorithmes d'apprentissage supervisé

Dans cette section, nous allons d'abord introduire la régression logistique, qui est un algorithme d'apprentissage linéaire, probabiliste et binomial. Nous montrerons ensuite comment nous pouvons la généraliser à plusieurs classes avec la régression logistique multinomiale. Ensuite, nous introduirons l'algorithme des Champs Conditionnels Aléatoires qui généralise la régression logistique multinomiale à l'apprentissage de séquences. Enfin, nous terminerons sur les Machines à Vecteurs de Support qui permettent d'apprendre des hypothèses non-linéaires.

B.2.1 La Régression Logistique

La régression logistique s'inscrit dans la tradition statistique de la régression linéaire[3], mais applique une fonction de seuil à $\theta^{\mathrm{T}}\mathbf{x}$ pour déterminer l'étiquette d'un exemple (Cox, 1959). Cette étiquette est soit positive, $y = 1$, soit négative, $y = -1$. Dans ce cadre, la régression logistique peut être considérée comme une modélisation d'une loi de Bernoulli.

Pour choisir une hypothèse $h_\theta(\mathbf{x})$ la régression logistique utilise un ensemble d'hypothèses ayant une forme de sigmoïde variant selon les paramètres :

$$\begin{aligned} h_\theta(\mathbf{x}) &= \mathrm{sigm}(\theta^{\mathrm{T}}\mathbf{x}) \\ &= \frac{1}{1+\exp(-\theta^{\mathrm{T}}\mathbf{x})} \end{aligned} \tag{eq.B.9}$$

[3] La régression linéaire prédit des valeurs continues, tandis que la régression logistique prédit des valeurs discrètes, mais en reposant sur un principe identique (Hastie *et al.*, 2009).

Cette fonction de seuil est appelée sigmoïde à cause de sa forme en S. Elle est définie sur l'intervalle [0,1]. Dans ce contexte, sa sortie est interprétée comme la probabilité d'appartenir à la classe positive. Cela implique attendre que $y = 1$ quand $\theta^{\mathrm{T}}\mathbf{x} \geq 0$ et $y = 0$ quand $\theta^{\mathrm{T}}\mathbf{x} < 0$, lorsque la régression logistique n'a pas de terme biais. Par exemple, pour un exemple où $\theta^{\mathrm{T}}\mathbf{x} = 1$, cet exemple aura une probabilité de 0,73 d'appartenir à la classe positive (Figure B.2).

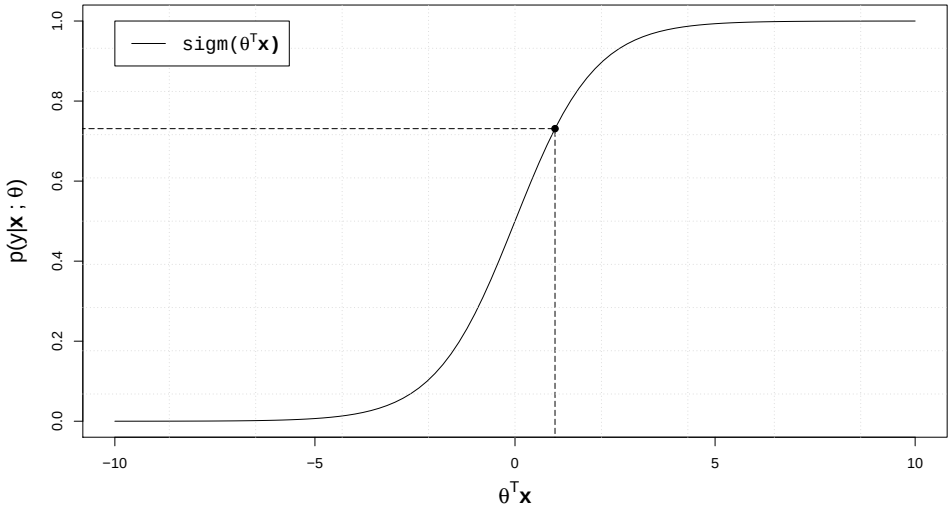

FIGURE B.2 : Fonction sigmoïde pour $\theta^T\mathbf{x} = 1$ sans terme biais

La fonction objective de la régression logistique est basée sur la fonction de vraisemblance (pour *likelihood*) (Cox et Snell, 1989). Celle-ci retourne la probabilité d'obtenir les exemples observés dans l'ensemble d'entraînement à partir de paramètres donnés. Ainsi, maximiser cette fonction permet d'obtenir les paramètres optimaux pour un ensemble d'apprentissage donné[4]. Dans ce contexte, l'objective est définie comme :

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) \qquad \text{(eq.B.10)}$$

La régression logistique suivant une loi de Bernoulli ($(y|x;\theta \sim Bernoulli(\phi))$), nous pouvons définir la fonction de masse suivante :

$$p(y|\mathbf{x}) = \begin{cases} h_\theta(\mathbf{x}), & \text{si } y = 1, \\ 1 - h_\theta(\mathbf{x}), & \text{si } y = 0. \end{cases} \qquad \text{(eq.B.11)}$$

[4] Notons que, bien que maximiser la fonction de vraisemblance semble intuitive, celle-ci reste néanmoins discutée comme outil d'inférence au sein de la communauté statistique, notamment concernant l'unicité des paramètres qu'elle peut retourner (Albert et Anderson, 1984).

Que nous réécrivons de manière plus compacte :

$$p(y|\mathbf{x}) = (h_\theta(\mathbf{x}))^y \ (1 - h_\theta(\mathbf{x}))^{1-y} \tag{eq.B.12}$$

Dans ce contexte, nous pouvons reprendre la fonction objective (eq.B.10) :

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) \\ &= \prod_{n=1}^{N} (h_\theta(\mathbf{x}_n))^{y_n} \ (1 - h_\theta(\mathbf{x}_n))^{1-y_n} \end{aligned} \tag{eq.B.13}$$

Cette fonction objective est passée à l'échelle logarithmique pour diminuer le coût de calcul et simplifier les dérivations[5] :

$$\begin{aligned} \ell(\theta) &= \log \mathcal{L}(\theta) \\ &= \log \left[\prod_{n=1}^{N} (h_\theta(\mathbf{x}_n))^{y_n} \ (1 - h_\theta(\mathbf{x}_n))^{1-y_n} \right] \\ &= \sum_{n=1}^{N} \left[y_n \log h_\theta(\mathbf{x}_n) + (1 - y_n) \log(1 - h_\theta(\mathbf{x}_n)) \right] \end{aligned} \tag{eq.B.14}$$

Le premier terme capture l'intuition que si $h(\mathbf{x}) \approx 0$ mais que $y = 1$, nous allons pénaliser l'algorithme par un coût large. Inversement, si $h(\mathbf{x}) \approx 1$ mais que $y = 0$, le deuxième terme de cette équation pénalisera également l'algorithme par un coût large. Dans ce contexte, nous allons chercher à maximiser cette fonction pour estimer les paramètres :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \left[y_n \log h_\theta(\mathbf{x}_n) + (1 - y_n) \log(1 - h_\theta(\mathbf{x}_n)) \right] \tag{eq.B.15}$$

Contrairement à la régression linéaire pour laquelle les résidus sont normalement distribués[6], il n'est pas possible d'obtenir une solution analytique pour estimer les paramètres de la régression logistique. Par contre, comme cette fonction est convexe (Figure B.3), il est possible d'utiliser une grande variété de processus itératifs tels que l'algorithme du gradient, la méthode du gradient conjugué (O'Leary P., 1996), l'algorithme BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) et son extension à mémoire limitée LM-BFGS (Liu et Nocedal, 1989). Toutes ces méthodes nécessitent initialement de pouvoir calculer le gradient de la fonction à optimiser.

[5] L'intérêt de la mettre à l'échelle logarithmique est que le logarithme est une fonction croissante convexe. Ainsi, plutôt que de maximiser $\mathcal{L}(\theta)$, il est plus simple de maximiser n'importe quelle fonction croissante de $\mathcal{L}(\theta)$.

[6] La régression linéaire suit une loi normale, c'est-à-dire $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$. (Williams, 1998)

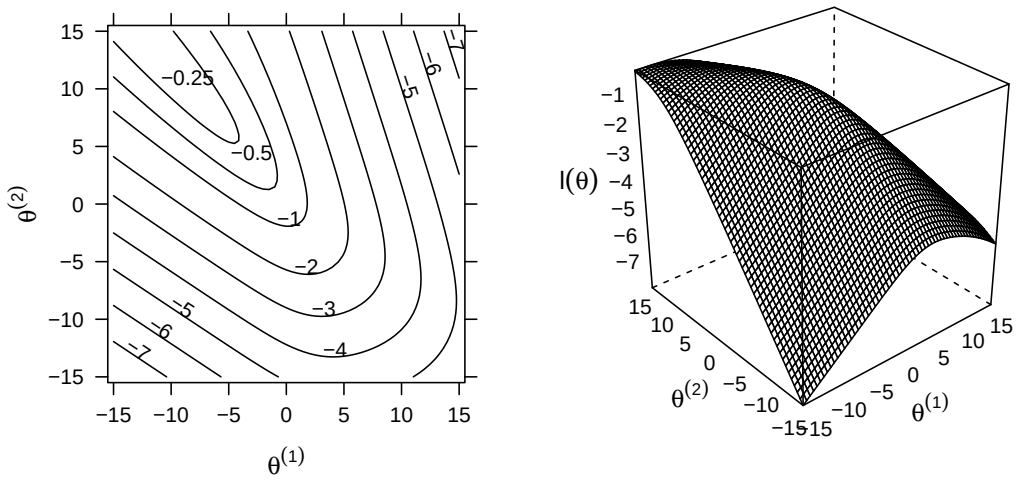

FIGURE B.3 : Exemple de log-vraisemblance pour une régression logistique à deux paramètres. À gauche : avec un plot de contours. À droite : avec une représentation tridimensionnelle. Les paramètres optimaux sont ceux qui maximisent cette fonction.

Le gradient de la régression logistique pour un paramètre $\theta^{(j)}$ donné est obtenu selon les règles usuelles de dérivation et permet de définir :

$$\begin{aligned}\frac{\partial}{\partial\theta^{(j)}}\ell(\theta) &= \frac{\partial}{\partial\theta^{(j)}}\sum_{n=1}^{N}\left[y_n \log h_\theta(\mathbf{x}_n) + (1-y_n)\log(1-h_\theta(\mathbf{x}_n))\right] \\ &= \sum_{n=1}^{N}\left[y_n x_n^{(j)} - h_\theta(\mathbf{x}_n)x_n^{(j)}\right]\end{aligned} \tag{eq.B.16}$$

Pour un paramètre fixé, cela revient à mesurer la différence entre les données observées et les sorties attendues du modèle selon ce paramètre. Dans la suite, nous donnons un exemple d'optimisation de l'objective par l'algorithme du gradient (algorithme 3). La règle de mise à jour des paramètres est :

$$\theta^{(j)} \leftarrow \theta^{(j)} + \alpha\frac{\partial}{\partial\theta^{(j)}}\ell(\theta) \tag{eq.B.17}$$

où α est le taux d'apprentissage. Il y a convergence dans l'optimisation de la fonction objective lorsque entre deux itérations t et $t-1$ nous avons (Malouf, 2002) :

$$\frac{|\ell(\theta_t) - \ell(\theta_{t-1})|}{\ell(\theta_t)} < \varepsilon \tag{eq.B.18}$$

où ε est la tolérance relative ayant, généralement, pour valeur 10^{-7}. Une fois la conver-

gence atteinte, les paramètres obtenus permettent la définition d'une frontière de décision en posant $\theta^{\mathrm{T}}\mathbf{x}= 0$. La figure B.4 donne un exemple de cette frontière de décision et des probabilités correspondantes pour un ensemble X de dimension 2 avec un ensemble de données non linéairement séparable.

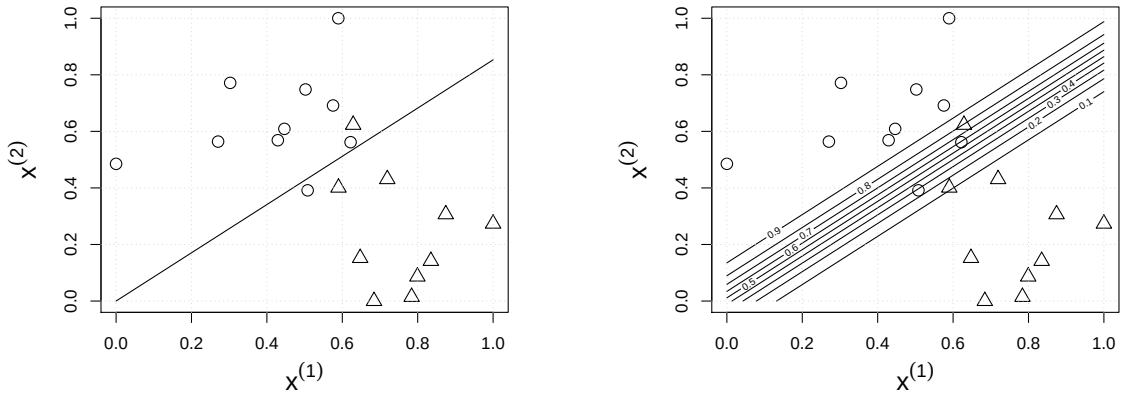

FIGURE B.4 : Exemple de frontière de décision et probabilités correspondantes pour une régression logistique avec deux paramètres

Algorithme 3 Algorithme du gradient pour l'optimisation de la fonction objective de la régression logistique

1: Initialisation du vecteur de paramètres $\theta \in \mathbb{R}^d$ à 0
2: Initialisation du vecteur gradient $\mathbf{g} \in \mathbb{R}^d$ à 0
3: **Tant Que** *convergence* n'est pas VRAI **Faire**
4: **Pour Chaque** j = 1, 2, ..., d **Faire**
5: $g^{(j)} \leftarrow \dfrac{\partial}{\partial\theta^{(j)}}\ell(\theta)$ `//Calcul du gradient`
6: **Fin Pour Chaque**
7: $\theta \leftarrow \theta + \alpha\ \mathbf{g}$ `//Mise à jour des paramètres`
8: **Si** aConvergé($\ell(\theta)$) :
9: $convergence \leftarrow$ VRAI
10: **Fin Si**
11: **Fin Tant Que**
12: Retourne le vecteur de paramètres θ

B.2.2 La Régression Logistique Multinomiale

La régression logistique présentée dans la section précédente est conçue pour les cas binomiaux. Or, il existe des situations où il est nécessaire de prendre en compte plusieurs classes. Une solution consiste à adapter les algorithmes de classification binomiale avec certaines stratégies telles que la *one-vs.-rest* ou la *one-vs.-one* (Bishop, 2006).

Cependant, ces stratégies ne permettent pas d'avoir des probabilités en sortie, car celles-ci se limitent généralement à (i) construire C-1 modèles et (ii) choisir la classe pour laquelle le score est maximisé. C'est pourquoi nous introduisons la régression logistique multinomiale qui permet la classification probabiliste multi-classes.

Selon le champ où cet algorithme d'apprentissage est utilisé, différents noms ont été proposés. Berger *et al.* (1996) utilisent le terme de classifieur d'entropie maximale et proposent une formulation duale liant la fonction du maximum de vraisemblance à la fonction d'entropie de Shannon (1948). Dans ce contexte, le principe d'entropie maximale vise à définir une contrainte pour chaque trait observé (Jaynes, 1957) et à choisir la distribution qui maximise l'entropie tout en restant consistante vis-à-vis de l'ensemble de ces contraintes.

D'autres auteurs tels que Malouf (2002) et Jansche (2005) utilisent le terme de modèle log-linéaire, qui reflète une caractéristique de la fonction objective.

En Statistiques, le terme de régression logistique multinomiale est davantage utilisé, car cet algorithme peut être considéré comme une généralisation à C classes de la régression logistique. Dans ce travail, nous utilisons ce dernier terme, car reflétant mieux les propriétés mathématiques de cet algorithme.

Pour choisir son hypothèse $h_\theta(\mathbf{x})$, la régression logistique multinomiale utilise un ensemble d'hypothèses où chaque h a la forme d'une fonction exponentielle normalisée, aussi appelée *softmax*[7]. Ainsi, en posant $y \in \{1, 2, \ldots, C\}$, nous obtenons pour une classe k donnée :

$$h_\theta(\mathbf{x})_k = \frac{\exp(\theta_k^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C} \exp(\theta_c^{\mathrm{T}}\mathbf{x})} \qquad \text{(eq.B.19)}$$

où le numérateur estime la possibilité que $\mathbf{x}$ appartienne à la classe $y = k$ et le dénominateur est une constante pour l'ensemble de la distribution.

[7] « The normalized exponential is also known as the softmax function, as it represents a smoothed version of the 'max' function because, if $a_k \gg a_j$ for all $j \neq k$, then $p(C_k|\mathbf{x}) \simeq 1$, and $p(C_j|\mathbf{x}) \simeq 0$. » (Bishop, 2006).

Ainsi, l'hypothèse généralisée $h_\theta(\mathbf{x})$ défini l'ensemble de la distribution et nous avons une matrice de paramètres où chaque colonne correspond à une classe :

$$h_\theta(\mathbf{x}) = \begin{bmatrix} \frac{\exp(\theta_1^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})} \\ \frac{\exp(\theta_2^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})} \\ \vdots \\ \frac{\exp(\theta_C^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})} \end{bmatrix}, \qquad \theta = \begin{bmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_C^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_C^{(2)} \\ \vdots & \vdots & \dots & \vdots \\ \theta_1^{(d)} & \theta_2^{(d)} & \dots & \theta_C^{(d)} \end{bmatrix} \tag{eq.B.20}$$

Dans ce cadre, nous pouvons définir la fonction de masse suivant une loi multinomiale, qui est une généralisation de la loi de Bernoulli :

$$p(y|\mathbf{x}) = \begin{cases} \frac{\exp(\theta_1^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})}, & \text{si } y = 1, \\ \frac{\exp(\theta_2^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})}, & \text{si } y = 2, \\ \vdots & \\ \frac{\exp(\theta_C^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})}, & \text{si } y = C, \end{cases} \tag{eq.B.21}$$

Que nous pouvons réécrire de manière plus compacte :

$$\begin{aligned} p(y|\mathbf{x}) &= \prod_{l=1}^{C}\left[\frac{\exp(\theta_l^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})}\right]^{1\{y=l\}} \\ &= \frac{\exp(\theta_y^{\mathrm{T}}\mathbf{x})}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x})} \end{aligned} \tag{eq.B.22}$$

Ainsi, nous obtenons la fonction de log-vraisemblance suivante :

$$\begin{aligned} \ell(\theta) &= \log\prod_{n=1}^{N} p(y_n|\mathbf{x}_n) \\ &= \sum_{n=1}^{N}\log p(y_n|\mathbf{x}_n) \\ &= \sum_{n=1}^{N}\log\left[\frac{\exp(\theta_{y_n}^{\mathrm{T}}\mathbf{x}_n)}{\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x}_n)}\right] \\ &= \sum_{n=1}^{N}\left[\theta_{y_n}^{\mathrm{T}}\mathbf{x}_n - \log\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x}_n)\right] \end{aligned} \tag{eq.B.23}$$

L'appellation de modèle log-linéaire provient de la dernière équation : à l'échelle logarithmique, la fonction de vraisemblance est linéaire dans le premier terme et le second terme est une constante indépendante du y donné en entrée.

De manière analogue à la régression logistique binomiale, le calcul de la dérivée partielle est effectué selon les règles usuelles de dérivation :

$$\begin{aligned}\frac{\partial}{\partial\theta_{y'}^{(j)}}\ell(\theta) &= \frac{\partial}{\partial\theta_{y'}^{(j)}}\sum_{n=1}^{N}\left[\theta_{y_n}^{\mathrm{T}}\mathbf{x}_n - \log\sum_{c=1}^{C}\exp(\theta_c^{\mathrm{T}}\mathbf{x}_n)\right] \\ &= \sum_{n=1}^{N}\left[[y_n = y']x_n^{(j)} - p(y'|\mathbf{x}_n)x_n^{(j)}\right]\end{aligned} \tag{eq.B.24}$$

Ainsi, nous pouvons observer que la régression logistique multinomiale est en effet une génération de la régression logistique binomiale à C classes.

De nombreux algorithmes d'optimisation ont été proposés en particulier pour la régression logistique multinomiale tels que le *Generalized Iterative Scaling* (Darroch et Ratcliff, 1972) et le *Improved Iterative Scaling* (Berger *et al.*, 1996). Les travaux de Malouf (2002) ont montré empiriquement que de meilleurs résultats étaient obtenus avec le LM-BFGS.

La figure B.5 donne un exemple[8] des frontières de décisions obtenues et des probabilités correspondantes pour un ensemble X de dimension 2 à trois classes non linéairement séparables.

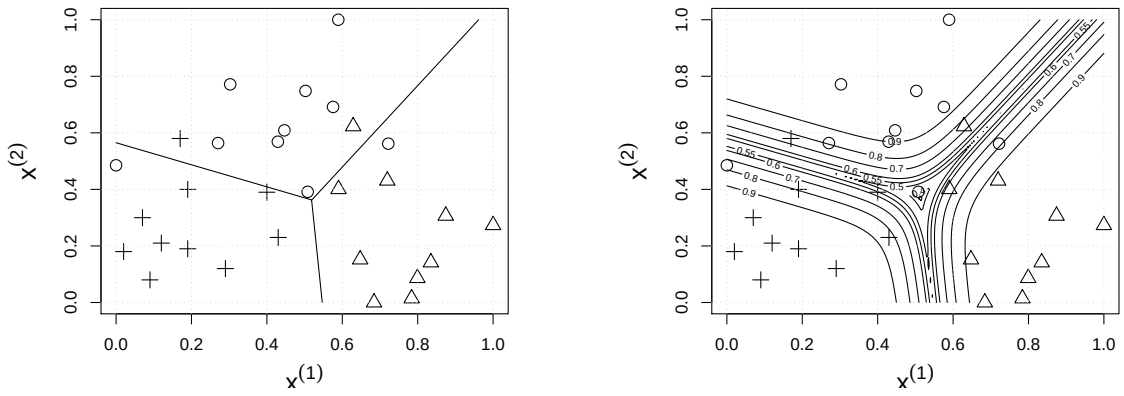

FIGURE B.5 : Frontière de décision et frontières avec probabilités pour une régression logistique multinomiale avec deux paramètres

[8] Le graphique a été construit avec notre implémentation R de la régression logistique multinomiale. Le code est sous licence libre et accessible à l'adresse : https://github.com/fauconnier/AMI.

B.2.3 Les Champs Conditionnels Aléatoires

Les Champs Conditionnels Aléatoires, aussi appelés *Conditional Random Fields* (CRF), sont des algorithmes d'apprentissage probabilistes introduits par Lafferty *et al.* (2001) qui permettent d'associer des séquences d'étiquettes à des séquences d'observations. Dans cette section, nous changeons légèrement la notation précédemment employée en définissant $\mathbf{y}$ comme une séquence d'étiquettes $y_1, y_2, \ldots, y_m$ et $\mathbf{x}$ comme une séquence d'observations $x_1, x_2, \ldots, x_m$. Nous limitons notre propos aux Champs Conditionnels Aléatoires de premier ordre.

Le principe général des CRF de premier ordre est d'établir que l'état de la i-ème position dépend uniquement de l'état de la position $(i-1)$:

$$\begin{aligned} p(y_1, y_2, \ldots, y_m | x_1, \ldots, x_m) &= \prod_{i=1}^{m} p(y_i | y_1, \ldots, y_{i-1}, x_1, \ldots, x_m) \\ &= \prod_{i=1}^{m} p(y_i | y_{i-1}, x_1, \ldots, x_m) \end{aligned} \tag{eq.B.25}$$

Dans ce cadre, pour modéliser la distribution multinomiale de probabilité sur une séquence $\mathbf{y}$, les CRF utilisent des hypothèses de forme comparable à la régression logistique multinomiale, mais en normalisant sur l'ensemble des séquences possibles des étiquettes. Ainsi pour une séquence fixée, les CRF prennent la forme :

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(\theta^{\mathrm{T}} F(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(\theta^{\mathrm{T}} F(\mathbf{x}, \mathbf{y}'))} \tag{eq.B.26}$$

Nous introduisons ici une notation avec fonction de traits[9] car elle simplifie le propos pour les CRF. Ici, la fonction $F(\mathbf{x}, \mathbf{y})$ est une *fonction globale* des traits qui retourne un vecteur de dimension d. Pour chaque dimension j, elle est définie :

$$F^{(j)}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} f^{(j)}(\mathbf{x}, y_{i-1}, y_i, i) \tag{eq.B.27}$$

Chaque dimension $F^{(j)}$ est calculée par la somme de la *fonction locale* $f^{(j)}$ appliquée sur les différentes transitions dans $y_1, \ldots, y_m$. Ces fonctions sont de deux types (Sha et Pereira, 2003) :

[9] Cette notation consiste à remplacer le vecteur de traits par une fonction de traits (*feature fonction*) qui retourne un vecteur qui sera non-nul pour une classe fixée. Citons Sutton et McCallum (2006) : « Rather than using one weight vector per class, (...), we can use a different notation in which a single set of weights is shared across all the classes. The trick is to define a set of feature functions that are nonzero only for a single class. To do this, the feature functions can be defined as $f_{y',j}(y, \mathbf{x}) = 1_{\{y'=y\}} x_j$ for the feature weights and $f_{y'}(y, \mathbf{x}) = 1_{\{y'=y\}}$ for the bias weights. Now we can use f_k to index each feature function $f_{y',j}$, and θ_k to index its corresponding weight $\theta_{y',j}$ ».

1. les traits d'états (*state features*) : ils portent uniquement sur un état donné à la position i dans la séquence et n'utilisent pas l'argument y_{i-1}
2. les traits de transitions (*transition features*) : ils portent sur un état donné à la position i dans la séquence et utilisent l'étiquette précédente y_{i-1}

De manière analogue à la régression logistique, les CRF utilisent une fonction de log-vraisemblance pour l'estimation de ses paramètres. Bien que l'hypothèse d'indépendance soit relâchée au sein d'une séquence, cette hypothèse reste maintenue entre les séquences[10]. Ainsi, la fonction objective des CRF est définie comme :

$$\begin{aligned}\ell(\theta) &= \sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{x}_n) \\ &= \sum_{n=1}^{N} \left[\theta^{\mathrm{T}} F(\mathbf{x}_n, \mathbf{y}_n) - \log \sum_{y' \in \mathcal{Y}^m} \exp(\theta^{\mathrm{T}} F(\mathbf{x}_n, \mathbf{y}')) \right]\end{aligned} \tag{eq.B.28}$$

La dérivée partielle de l'objective se calcule avec les règles usuelles. Ceci nous permet de poser :

$$\frac{\partial}{\partial \theta^{(j)}} \ell(\theta) = \sum_{n=1}^{N} F^{(j)}(\mathbf{x}_n, \mathbf{y}_n) - \sum_{i=1}^{N} \sum_{\mathbf{y}' \in \mathcal{Y}^m} p(\mathbf{y}'|\mathbf{x}_n) F^{(j)}(\mathbf{x}_n, \mathbf{y}') \tag{eq.B.29}$$

Sur le plan pratique, deux difficultés interviennent dans les CRF. La première se situe au niveau de l'activation (eq.B.26), et la seconde réside dans le calcul du gradient (eq.B.29). Leur cause commune est le calcul de la constante de normalisation :

$$\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(\theta^{\mathrm{T}} F(\mathbf{x}, \mathbf{y}')) \tag{eq.B.30}$$

Cette constante est incalculable en des temps raisonnables car elle implique une somme de tout l'ensemble $\mathcal{Y}^m$ de taille r^m où m est la longueur de la séquence et $r = |\mathcal{Y}|$ est le nombre d'étiquettes possibles. Pour contrer cette difficulté, les CRF utilisent des stratégies de programmation dynamique identiques à celles déjà utilisées dans le Modèle de Markov Caché (*Hidden Markov Model*) (Rabiner, 1989).

Pour la fonction d'activation, un algorithme de Viterbi permet d'obtenir, pour une séquence $\mathbf{x}$ donnée, la séquence d'étiquettes la plus probable. Cela revient à trouver la séquence qui maximise les scores pour chacune de ses transitions y_{i-1}, y_i.

[10] Il est dit que les exemples/séquences sont indépendants et identiquement distribués (iid).

Cela nécessite de simplifier et maximiser la fonction d'activation :

$$\begin{aligned}
\underset{\mathbf{y}\in\mathcal{Y}^m}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}) &= \underset{\mathbf{y}\in\mathcal{Y}^m}{\operatorname{argmax}} \quad \frac{\exp(\theta^{\mathrm{T}} F(\mathbf{x},\mathbf{y}))}{\sum_{\mathbf{y}'\in\mathcal{Y}^m}\exp(\theta^{\mathrm{T}} F(\mathbf{x},\mathbf{y}'))} \\
&= \underset{\mathbf{y}\in\mathcal{Y}^m}{\operatorname{argmax}} \quad \theta^{\mathrm{T}} F(\mathbf{x},\mathbf{y}) \\
&= \underset{\mathbf{y}\in\mathcal{Y}^m}{\operatorname{argmax}} \quad \sum_{i=1}^{m}\sum_{j=1}^{d}\theta^{(j)} f^{(j)}(\mathbf{x}, y_{i-1}, y_i, i) \\
&= \underset{\mathbf{y}\in\mathcal{Y}^m}{\operatorname{argmax}} \quad \sum_{i=1}^{m}\theta^{\mathrm{T}}\mathbf{f}(\mathbf{x}, y_{i-1}, y_i, i)
\end{aligned} \tag{eq.B.31}$$

Le calcul de la constante pour le gradient repose aussi sur un algorithme de programmation dynamique. Généralement, l'algorithme de *forward-backward* est utilisé et permet l'entraînement d'un modèle en un temps raisonnable (Sutton et McCallum, 2006).

B.2.4 Les Machines à Vecteurs de Support

La régression logistique et sa généralisation à plusieurs classes sont des algorithmes d'apprentissages probabilistes puissants et utiles pour les cas où il est possible d'apprendre des hypothèses linéaires à partir des données. Cependant, ces algorithmes ne sont pas capables d'apprendre des hypothèses non linéaires qui peuvent apparaître dans certains cas d'apprentissage.

Dans cette section, nous introduisons les Machines à Vecteurs de Support (SVM) qui sont des algorithmes d'apprentissage binaires capables d'apprendre des hypothèses non linéaires.

Deux principes fondateurs se trouvent derrière le SVM : (i) il s'agit de séparer les données par un hyperplan à marge maximale et (ii) l'utilisation de méthodes noyaux permet de transformer l'espace de traits en un espace de plus grande dimension, voire infinie, afin d'apprendre des hypothèses non linéaires.
Ces deux éléments ont fait l'objet de travaux précédents : Vapnik et Lerner (1963) ont montré l'intérêt d'un classifieur à vaste marge et Aizerman *et al.* (1964) ont souligné l'intérêt des fonctions noyaux dans le contexte de l'apprentissage. En 1992, ces deux notions sont rassemblées pour former le SVM (Boser *et al.*, 1992). Le principe d'une marge souple permettant le traitement des données extrêmes (*outliers*) sera proposée en 1995 (Cortes et Vapnik, 1995).

Dans sa version linéaire, le SVM apprend des hypothèses ayant une forme comparable à celles du perceptron (Rosenblatt, 1958) :

$$h_{\theta,b}(\mathbf{x}) = sgn(\theta^{\mathrm{T}}\mathbf{x} + b) \tag{eq.B.32}$$

où b est le terme biais[11] et θ sont les paramètres. Dans ce cadre, $y \in \{-1, 1\}$ et la fonction sgn permet de choisir la classe d'appartenance pour un vecteur $\mathbf{x}$ en fonction du signe de la valeur issue du produit salaire. Pour estimer le vecteur de paramètres θ, le SVM cherche l'hyperplan unique qui sépare les deux classes avec des marges maximales.

Deux types de marges sont utilisés dans le SVM : (i) les marges fonctionnelles, et (ii) les marges géométriques.

Marge fonctionnelle Soit un exemple $(\mathbf{x}_n, y_n)$, nous définissons la marge fonctionnelle pour les paramètres θ et b fixés comme suit :

$$m_n = y_n(\theta^{\mathrm{T}}\mathbf{x}_n + b) \tag{eq.B.33}$$

Ainsi, la marge fonctionnelle peut être vue comme la distance entre un exemple et un hyperplan défini par $\theta^{\mathrm{T}}\mathbf{x}= 0$. La multiplication par le terme y_n permet d'indiquer si l'exemple est correctement classé. Si $m_n > 0$ alors l'exemple n est correctement prédit. Si $m_n = 0$ alors l'exemple est situé sur l'hyperplan.

En généralisant à un ensemble d'entraînement $\mathcal{D}$ linéairement séparable, nous définissons la marge fonctionnelle générale comme la marge fonctionnelle la plus petite des exemples d'entraînement pris individuellement :

$$m = \min_{i=1,\ldots,N} m_n \tag{eq.B.34}$$

Ainsi, maximiser les marges sous la contrainte que chacun des exemples est correctement classé revient au problème d'optimisation suivant :

$$\begin{aligned} \max_{\theta,b} \quad & m \\ \text{s. t.} \quad & y_n(\theta^{\mathrm{T}}\mathbf{x}_n + b) \geq m, \qquad n = 1, \ldots, N. \end{aligned} \tag{eq.B.35}$$

Marge géométrique En l'état, les paramètres θ et b peuvent être arbitrairement ajustés pour maximiser la valeur de m[12]. Pour corriger cela, la marge géométrique normalise m par la norme de θ. Le problème d'optimisation est réécrit :

$$\begin{aligned} \max_{\theta,b} \quad & \frac{m}{||\theta||} \\ \text{s. t.} \quad & y_n(\theta^{\mathrm{T}}\mathbf{x}_n + b) \geq m, \qquad n = 1, \ldots, N. \end{aligned} \tag{eq.B.36}$$

[11] Dans cette section, nous traitons différemment le terme biais en le sortant du produit scalaire. Ainsi, ici, b joue le rôle de $\theta^{(0)}$ et θ est le vecteur composé de $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(d)}$. Ceci permet de faciliter le propos dans la suite de la section.

[12] Il est possible d'augmenter arbitrairement les marges fonctionnelles en faisant par exemple $(5\theta, 5b)$ pour obtenir un résultat cinq fois supérieur. Or, l'hypothèse $h_{\theta,b}(\mathbf{x})$ ne réagit qu'au signe et non à la magnitude du produit scalaire.

Cette normalisation implique que la marge géométrique est insensible à l'augmentation des valeurs. Dans ce contexte, nous posons la contrainte suivante :

$$m = 1 \tag{eq.B.37}$$

Maximiser $1/||\theta||$ étant équivalent à minimiser $||\theta||^2$, nous obtenons le problème d'optimisation primal du SVM (Boser *et al.*, 1992) :

$$\begin{aligned} \min_{\theta,b} \quad & \frac{1}{2}||\theta||^2 \\ \text{s. t.} \quad & y_n(\theta^{\mathrm{T}}\mathbf{x}_n + b) \geq 1, \qquad n = 1, \ldots, N. \end{aligned} \tag{eq.B.38}$$

Le facteur $1/2$ n'est utile ici que pour des simplifications mathématiques[13]. Les exemples pour lesquels $y_n(\theta^{\mathrm{T}}\mathbf{x}_n + b) = 1$ sont les vecteurs de support, c'est-à-dire les exemples présents sur les marges.

Apprendre les marges et les vecteurs de support est un problème d'optimisation quadratique contraint. En appliquant la méthode des multiplicateurs de Lagrange (Luenberger, 1984), nous obtenons la formulation duale du SVM :

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \\ \text{s. t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, l. \\ & \sum_{i=1}^{l} \alpha_i y_i = 0. \end{aligned} \tag{eq.B.39}$$

où α_i avec $i, \ldots, l$ sont les multiplicateurs de Lagrange associés au l supports de vecteurs. Cette formulation duale du SVM peut être optimisée avec des algorithmes tels que le *Sequential Minimal Optimization* (Platt *et al.*, 1998).

Une fois entraîné, la relation entre la formulation primale et duale est donnée par l'équivalence suivante :

$$\theta = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i \tag{eq.B.40}$$

Ainsi, en utilisant cette équivalence dans la fonction d'activation du SVM (eq.B.32), nous pouvons l'exprimer avec les vecteurs de support. Il s'agit de la formulation duale de l'activation du SVM :

$$h_\alpha(\mathbf{x}) = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b \tag{eq.B.41}$$

[13] « The factor one half has been included for cosmetic reasons; it does not change the solution. » (Boser *et al.*, 1992).

La figure Figure B.6 donne un exemple[14] de l'application de cette fonction d'inférence après entraînement sur un ensemble de dimension 2. L'hyperplan séparateur a une distance fonctionnelle de 1 par rapport aux deux marges. Trois vecteurs de supports constituent ce modèle.

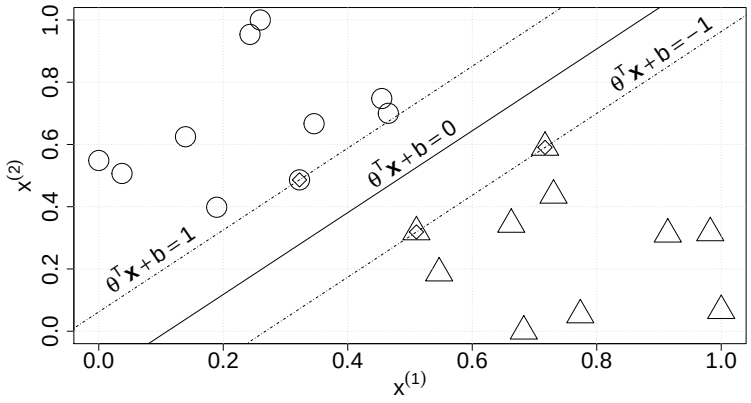

FIGURE B.6 : Exemples répartis selon 2 classes et séparés par un hyperplan à marges maximales

Méthodes à noyaux L'expression du SVM par le produit scalaire entre les vecteurs de support et un $\mathbf{x}$ donné permet l'utilisation de méthodes noyaux pour apprendre des fonctions non linéaires. Le principe d'un noyau est de transformer l'espace de traits original dans un autre de plus grande dimension où il aura plus de chance d'être séparé linéairement :

> A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated. (Kim *et al.*, 2005)

L'intégration des noyaux dans le SVM revient à remplacer tous les produits scalaires par une fonction noyau choisie. Par conséquent, la formulation duale de la fonction d'activation du SVM deviendra :

$$h_\alpha(\mathbf{x}) = \sum_{i=1}^{l} y_i \alpha_i \; K(\mathbf{x}_i, \mathbf{x}) + b \qquad \text{(eq.B.42)}$$

où K est la fonction noyau qui prend en entrée deux vecteurs. Cette fonction est équivalente au produit scalaire suivant :

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}')^{\mathrm{T}} \phi(\mathbf{z}) \qquad \text{(eq.B.43)}$$

[14] Exemple réalisé avec l'implémentation LibSVM (Chang et Lin, 2013).

où ϕ est une fonction de transformation. Par exemple, dans la figure B.7[15] l'application de la transformation $\phi(\mathbf{x}) = \mathbf{x}^2$ permet, dans le nouvel espace de plus grande dimension, de séparer linéairement les deux classes.

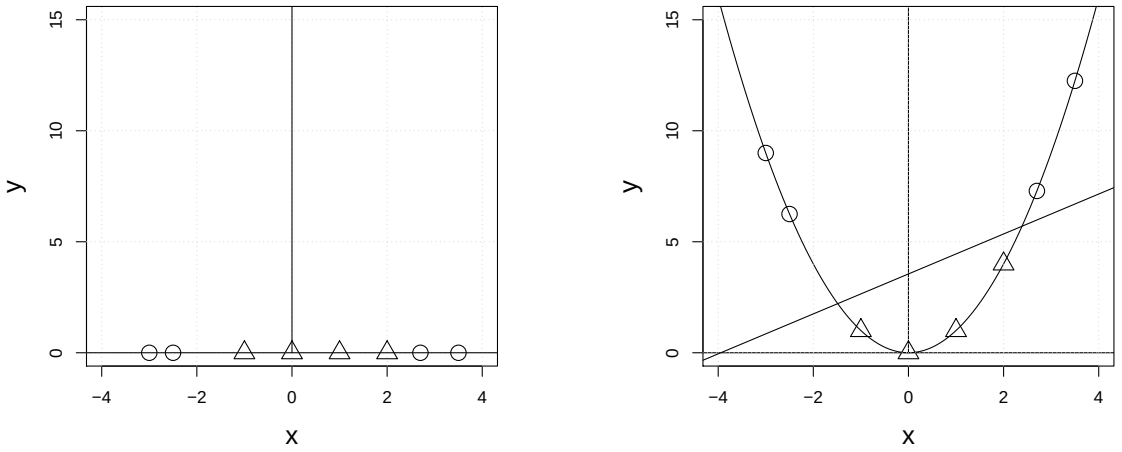

FIGURE B.7 : Application d'une fonction de transformation $\phi(x) = x^2$ à un ensemble de deux classes dans un espace de dimension 1 pour permettre leur séparation linéaire dans un espace de dimension 2

L'intérêt d'utiliser une fonction noyau est qu'il n'est pas obligatoire d'appeler explicitement la transformation ϕ, potentiellement coûteuse (Ben-Hur et Weston, 2010)[16]. Notons que seules les fonctions répondant aux contraintes posées par le théorème de Mercer sont utilisables comme fonction noyau au sein d'un SVM.

Dans cette thèse, nous utilisons un noyau gaussien pour apprendre des hypothèses non-linéaires. Celui désigne une fonction de base radiale (*radial basis function*) qui a la forme :

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{-||x - z||^2}{2\sigma^2}\right) \qquad \text{(eq.B.44)}$$

où σ définit la sensibilité de la courbe gaussienne[17]. Cette fonction noyau peut être considérée comme une mesure de similarité entre deux vecteurs. La figure B.8 montre un exemple simplifiée pour uniquement deux valeurs x et y entre $[-5, 5]$.

[15] Exemple adapté de (Urieli, 2013).

[16] Par exemple, pour le noyau polynomial d'ordre 2, une transformation $\phi(\mathbf{x})$ où $\mathbf{x}$ est de dimension 2 ($d = 2$) donnera :

$$\phi(\mathbf{x}) = x^{(1)}x^{(1)}, x^{(1)}x^{(2)}, x^{(2)}x^{(1)}, x^{(2)}x^{(2)}, \sqrt{2cx^{(1)}}, \sqrt{2cx^{(2)}}, c$$

Cela aura une complexité $O(d^2)$. *A contrario*, la fonction noyau $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z} + c)^2$, aura une complexité de $O(d)$ pour un résultat identique au produit scalaire $\phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{z})$.

[17] Notons que le facteur $\frac{1}{2\sigma^2}$ est aussi noté γ.

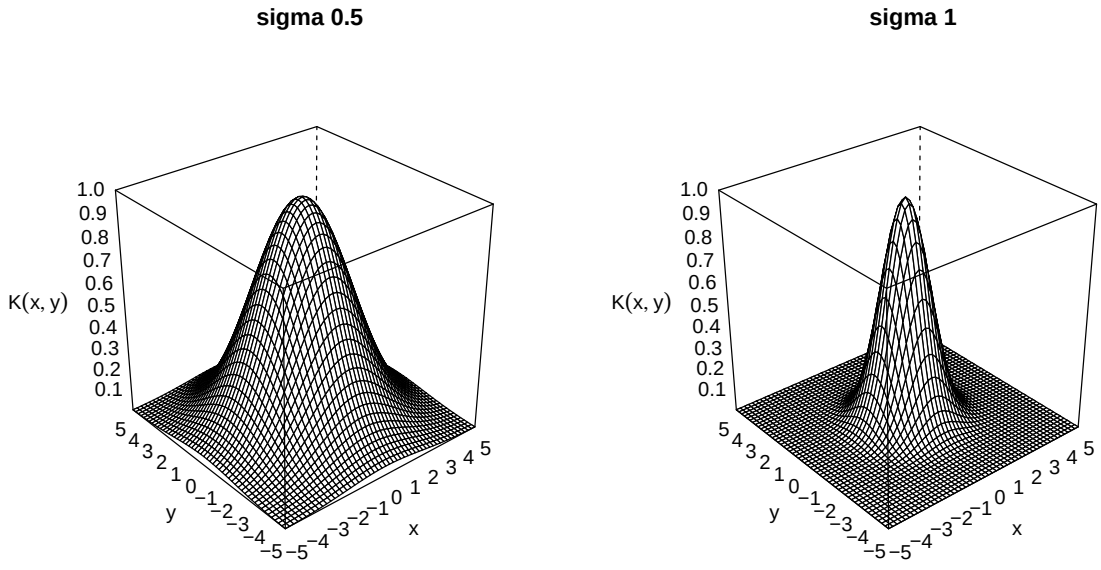

FIGURE B.8 : Exemples d'application d'une fonction gaussienne pour deux valeurs x et y entre $[-5, 5]$ avec un σ 0,5 et un σ à 1

B.3 Comparaison entre les algorithmes

Dans cette section, nous apportons un regard qualitatif sur les algorithmes présentés dans cette annexe. Le propos est scindé en deux parties : (i) nous comparons les algorithmes de classification non-séquentiels en opposant les modèles linaires probabilistes et les SVM, ensuite (ii) nous comparons le CRF aux autres algorithmes d'apprentissage séquentiels.

Comparaison entre les modèles linéaires et les SVM Le choix d'un modèle linéaire probabiliste ou d'un SVM dépend de plusieurs facteurs. Trois points peuvent être considérés dans leur comparaison :

- Premièrement, la régression logistique et son extension multinomiale sont adaptés aux cas où il n'existe pas parmi les traits des prédicteurs suffisamment forts pour déterminer avec précision la classe d'appartenance d'un exemple donné. Dans ce cas, une estimation probabiliste est généralement préférée. C'est souvent le cas lorsqu'il y a beaucoup de bruit dans les données. Le SVM sera préférable dans les cas où il y a un très grand nombre de dimensions mais qu'il existe des prédicteurs qui peuvent déterminer avec une grande certitude l'appartenance à une classe et permettre la maximisation de la marge (Vapnik, 1995)[18]. Notons qu'il existe des

[18] Cela peut être le cas, par exemple, dans la reconnaissance de chiffres où certains motifs sont exprimés

méthodes pour obtenir une sortie probabilisée d'un SVM binaire, telles que le *Platt scaling* (Platt, 1999), mais qu'elles restent néanmoins non triviales en nécessitant généralement l'entraînement additionnel d'une régression logistique.

- Deuxièmement, les modèles linéaires probabilistes considèrent tous les exemples lors de l'entraînement et de l'inférence, tandis que le SVM construira son modèle sur la base unique des vecteurs de support. Ajouter davantage de données pour l'entraînement d'un SVM ne conduit pas systématiquement à une amélioration des résultats, si les nouveaux exemples sont loin des marges. En fonction du problème et du nombre de données à disposition, l'une ou l'autre des classes d'algorithmes doit être préférée.

- Troisièmement, le SVM permet d'apprendre des fonctions non linéaires. Cet avantage est utile, mais augmente le risque de sur-apprentissage, c'est-à-dire de non généralisation sur de nouvelles données (Cawley et Talbot, 2010). Bien que le SVM intègre dans son extension à marge souple (Cortes et Vapnik, 1995) un type de régularisation avec sa constante C, il reste nécessaire de paramétrer finement les hyperparamètres du SVM pour éviter le sur-apprentissage. Tâche qui peut s'avérer coûteuse car nécessitant une recherche systématique, aussi appelée *grid search* (Chang et Lin, 2013). Les modèles probabilistes linéaires ont souvent pour seuls paramètres le nombre d'itérations alloués à l'optimisation de l'objective et un paramètre de seuillage qui permet une *régularisation* naïve mais souvent efficiente.

	R. L. binomiale	R. L. multinomiale	SVM
Complexité	simple	moyenne	grande
Entraînement	rapide	moyen	lent
Paradigme	binomial	multinomial	binaire
Probabilité	support	support	non-support
Hypothèse	linéaire	linéaire	linéaire et non linéaire
Hyperparamètres	itérations, seuil	itérations, seuil	constante C, σ, seuil

TABLE B.1 : Comparaison entre les algorithmes de classification non-séquentielle

La table B.1 synthétise les principales propriétés de chacun des algorithmes comparés ici. Notons que tous respectent une certaine forme du rasoir d'Occam : c'est-à-dire que ces algorithmes auront tendance à pénaliser un trait qui co-occurre rarement avec une classe plutôt qu'un trait qui n'apparaît jamais avec cette classe (Urieli, 2013).

Algorithmes d'apprentissage séquentiels Dans ce travail de thèse, nous avons choisi d'utiliser des Champs Conditionnels Aléatoires (CRF). D'autres algorithmes d'apprentissage séquentiels existent, néanmoins les CRF montrent des propriétés intéressantes.

dans un grand nombre de dimensions mais restent déterminants.

- Premièrement, contrairement aux modèles génératifs, tel que les modèles de Markov Caché (*Hidden Markov Model* pour HMM) (Rabiner, 1989), l'intégration de traits est plus simple pour les CRF et il est possible de modéliser des dépendances complexes (Lafferty *et al.*, 2001).

- Deuxièmement, contrairement aux Modèles Markoviens d'Entropie Maximale (*Maximum Entropy Markov Models* pour MEMM) (McCallum *et al.*, 2000), les CRF permettent de régler le problème du bais en normalisant sur l'ensemble de la séquence et non sur chacun des états. Lafferty *et al.* expliquent :

  > The Markovian assumptions in MEMMs and similar state-conditional models insulate decisions at one state from future decisions in a way that does not match the actual dependencies between consecutive states.

- Enfin, les CRF bénéficient des méthodes de programmations dynamiques héritées des précédents modèles séquentiels. Celles-ci font des CRF un algorithme efficient pour l'apprentissage de séquences, comme l'ont montré empiriquement les travaux sur le parsing (Sha et Pereira, 2003), l'extraction d'information (Sarawagi et Cohen, 2004) ou encore la segmentation en tokens (Tseng *et al.*, 2005).

Nuançons néanmoins le propos en notant que des études théoriques, telles que le « *no free lunch* » théorème (Wolpert et Macready, 1997), ou empiriques (Daelemans et Hoste, 2002), ont montré qu'il n'existait pas d'algorithmes d'apprentissage universellement meilleurs que les autres. Seules certaines classes d'algorithmes d'apprentissage et certains types d'optimisation montrent, empiriquement, des performances supérieures sous des conditions données.

Annexe C

Annexes pour les structures énumératives

Sommaire

C.1 Algorithme d'alignement positionnel

Soit L, $L_{initial}$, L^{-} des listes d'alignements unitaires d'annotation, $L_{initial}$ est triée, et $\dot{a}$ un alignement unitaire donné (Mathet et Widlöcher, 2011).

Algorithme 4 Algorithme de Mathet et Widlöcher (2011) pour une solution approchée d'une configuration idéale d'alignements

1: $L \leftarrow L_{initial}$
2: $i \leftarrow 0$
3: **Tant Que** $i < taille(L) - 1$ **Faire**
4: $\quad \dot{a} \leftarrow L[i]$
5: $\quad L^{-} \leftarrow L[i+1, (taille(L) - 1)]$
6: $\quad$ Retirer de L^{-} les alignements contenant une unité de $\dot{a}$
7: $\quad L \leftarrow L^{-}$
8: $\quad i \leftarrow i + 1$
9: **end Tant Que**

C.2 Interface pour la correction des alignements positionnels

Nous donnons des captures d'écran de l'interface développée pour la correction manuelle des alignements positionnels entre les structures énumératives. Les figures C.1 et C.2 donnent des exemples d'alignements qui ne nécessitent pas de correction manuelle. La figure C.3 montre une erreur d'alignement.

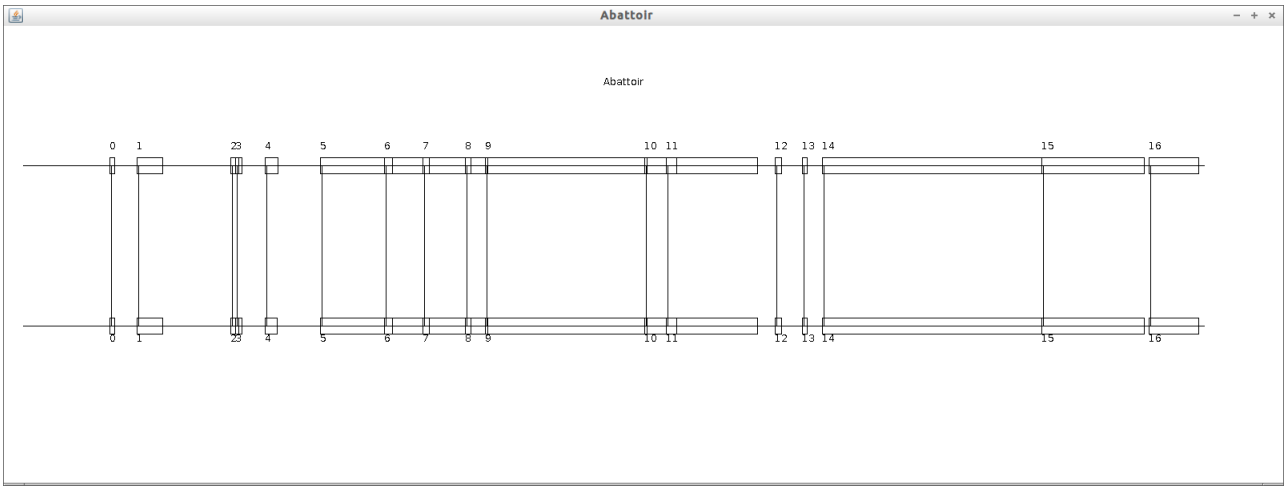

FIGURE C.1 : Interface pour la correction manuelle des alignements dans le document abattoir

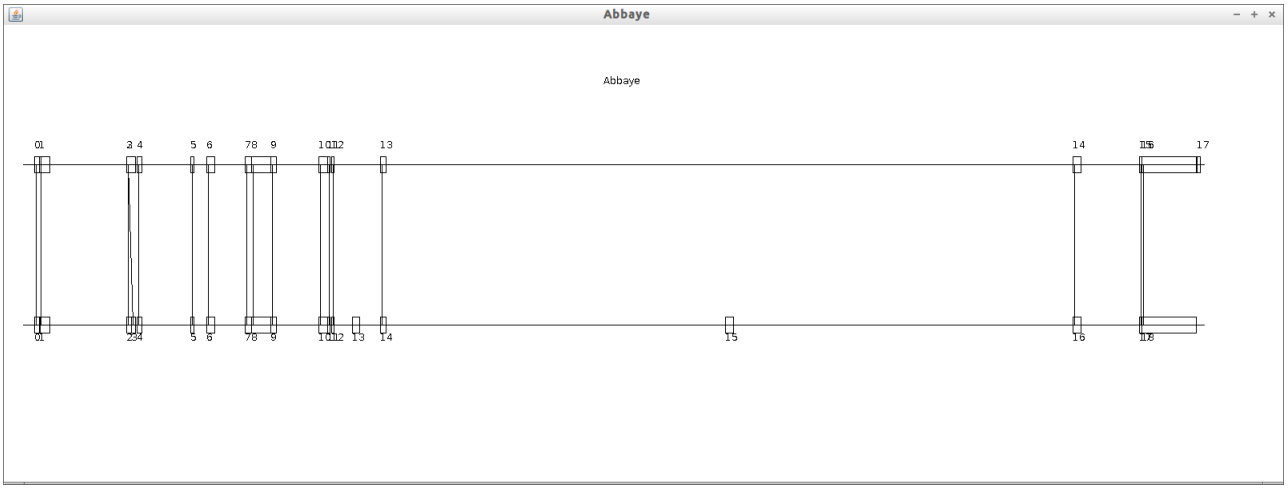

FIGURE C.2 : Interface pour la correction manuelle des alignements dans le document abbaye

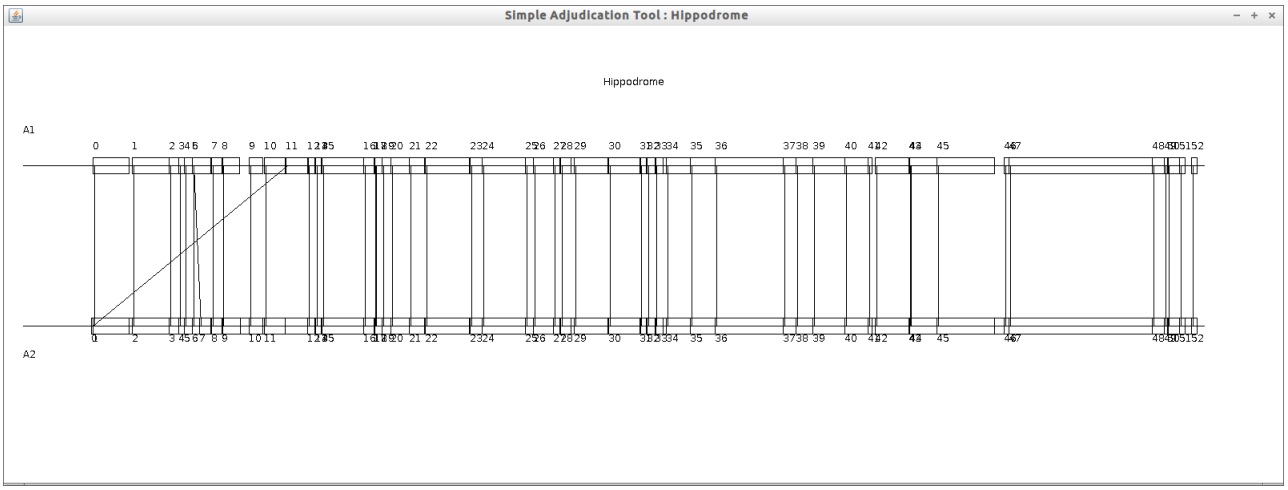

FIGURE C.3 : Interface pour la correction manuelle des alignements dans le document hippodrome

C.3 Tableau d'alignement des annotations visuelles

Dans ce tableau, nous reportons, pour chaque document (169 au total), les nombres des annotations effectuées par les deux annotateurs étudiants pour la phase d'annotation visuelle (Section 6.2.2). L'union et l'intersection sont définies avec l'aide de l'algorithme de recherche d'alignement de Mathet et Widlöcher (2011) (Annexe C.1). La solution obtenue est validée manuellement avec une interface de correction développée (Annexe C.2).

document	Étudiant 1	Étudiant 2	union	intersection	validation
Abattoir	17	17	17	17	17
Abbaye	18	19	20	17	17
Amer	1	1	1	1	1
Anse	2	2	2	2	2
Antenne	4	4	4	4	4
Aquaculture	22	27	27	22	22
Aqueduc	4	4	4	4	4
Arbre	34	29	29	25	24
Arc_de_triomphe	14	16	16	14	14
Atelier	5	7	8	5	4
Autoroute	22	28	28	20	20
Aven	7	14	14	7	7
Baie	2	3	3	2	2
Balise	5	4	5	4	4
Banc	5	6	6	5	5
Barrage	39	43	43	36	35
Base	8	8	8	8	7
Beffroi	2	2	3	1	1
Bief	6	13	13	6	6
Bois	55	64	64	50	49
Borne	6	7	7	6	6
Bosquet	5	6	6	5	5
Bouchot	1	1	1	1	1
Bureau_de_poste	1	1	1	1	1
Cabane	9	11	11	8	7
Calanque	4	4	4	4	4
Camp	9	10	10	9	9
Camping	23	22	23	20	20
Canal	6	8	8	6	6
Canalisation	4	4	4	4	4
Canton	2	5	5	2	2
Cap	5	7	7	5	5
Carrefour	1	4	4	1	1

Casemate	3	3	3	3	2
Caserne_de_pompiers	2	2	2	2	2
Centrale_thermique	20	23	23	20	19
Centre_commercial	7	9	9	6	5
Centre_culturel	3	3	3	3	3
Chalet	6	7	7	5	5
Champ_de_tir	2	3	3	2	2
Chaos	6	7	7	6	6
Chapelle	3	3	3	3	3
Chemin	2	2	3	2	1
Chute_d'eau	1	1	1	1	1
Cirque	53	60	60	49	48
Citadelle	8	9	9	8	8
Clocher	5	5	5	3	3
Cluse	2	3	3	2	2
Colline	2	2	2	2	2
Colonie_de_vacances	1	1	1	1	1
Construction	15	12	12	12	12
Cours_d'eau	8	8	8	7	7
Crevasse	0	1	1	0	0
Crique	2	2	2	2	2
Culte	6	5	5	5	5
Culte_protestant	7	6	6	6	6
Culture	39	35	35	31	31
Delta	3	6	6	3	3
Digue	15	17	17	13	13
Doline	5	4	4	4	4
Dolmen	6	6	7	5	5
Donjon	2	1	1	1	1
Dune	4	5	5	4	4
Eau	37	38	38	35	35
Enseignement	10	7	7	7	7
Enseignement_primaire	1	1	1	1	1
Enseignement_secondaire	10	10	10	10	10
Escalier	2	13	13	2	2
Estuaire	26	25	25	22	21
Fabrique	6	6	6	6	5
Falaise	9	9	9	9	9
Faubourg	6	6	6	5	5
Fleuve	8	8	9	8	7
Foire	3	3	3	3	3
Fontaine	15	14	14	13	13
Funiculaire	3	4	4	2	2

Gare	5	7	7	5	5
Gazoduc	7	7	7	5	5
Gendarmerie	4	3	4	2	2
Glacier	2	41	41	2	2
Golf	49	41	41	32	32
Gouffre	3	4	4	3	3
Grand_magasin	11	9	9	8	8
Grotte	3	3	3	3	3
Haie	9	10	10	7	7
Halle	4	5	5	4	4
Hameau	5	3	3	2	2
Haras	0	1	1	0	0
Haut_fourneau	18	16	16	15	15
Hippodrome	53	53	53	53	52
Hospice	3	4	4	3	3
Institut	1	1	1	1	1
Isthme	2	2	2	2	2
Jardin	16	15	16	13	12
Lac	12	15	16	11	11
Lieu-dit	4	4	4	4	4
Lotissement	5	8	8	5	4
Mangrove	9	10	12	9	9
Manufacture	2	4	4	2	2
Marais_salant	3	6	7	3	2
Mare	35	35	37	29	27
Menhir	10	13	13	9	9
Mer	8	11	11	7	7
Militaire	6	6	7	5	5
Minaret	4	5	5	4	4
Mine	7	6	6	6	5
Minoterie	3	4	4	3	3
Mont	7	8	8	7	7
Montagne	56	44	44	36	36
Monument	4	4	5	3	3
Moraine	2	2	2	2	2
Mur_anti-bruit	7	5	5	5	5
Observatoire	1	1	1	1	1
Oronyme	20	1	17	1	1
Parc	9	3	5	3	3
Parc_de_loisirs	2	3	3	2	2
Parc_national	0	2	2	0	0
Parc_zoologique	2	2	2	2	2
Parking	7	7	8	6	6

Passage	3	3	3	3	3
Passe	5	12	12	5	5
Phare	13	13	13	13	13
Pic	6	7	7	6	6
Piscine	8	8	9	8	8
Place	2	3	3	2	2
Plaine	2	2	2	2	2
Plantation	1	1	1	1	1
Plateau	7	9	9	6	6
Point_de_vue	2	3	3	2	2
Pointe	2	6	6	2	2
Pont	3	3	3	3	3
Pont-canal	1	1	1	1	1
Pont_mobile	1	1	1	1	1
Pont_suspendu	6	6	6	6	6
Pont_transbordeur	1	1	1	1	1
Port	12	11	11	11	11
Porte_de_ville	12	12	12	12	12
Portique	2	2	2	2	2
Prison	4	4	4	4	4
Puits	4	4	4	4	4
Radar	16	16	16	15	15
Radier	4	4	4	4	4
Rigole	2	2	2	2	2
Rocher	5	5	5	5	5
Route	5	5	5	5	5
Saline	2	3	3	2	2
Sanatorium	1	1	1	1	1
Science	18	19	19	18	18
Scierie	1	1	1	1	1
Sentier	4	4	4	4	4
Silo	3	3	3	3	3
Source	9	9	9	9	9
Sport	13	12	13	12	12
Station	5	5	6	5	5
Synagogue	3	3	4	3	3
Talus	1	2	3	1	1
Temple	5	5	6	5	5
Terril	1	1	1	1	1
Tour	4	4	4	4	3
Tribunal	2	2	2	2	2
Tumulus	2	3	3	2	2
Tunnel	1	1	2	1	1

Usine	5	5	5	5	5
Val	6	6	6	6	6
Vanne	6	5	6	5	5
Verger	2	2	2	2	2
Viaduc	3	4	4	3	3
Vigne	27	33	33	27	27
Volcan	11	10	10	10	10
Totaux	1406	1517	1562	1239	1217

TABLE C.1 : Table des alignements pour la phase visuelle d'annotation

C.4 Analyse des traits pour la tâche T_Onto

Le tableau ci-dessous ordonne les 10 traits présentant les plus grandes valeurs absolues de corrélation à la présence d'une relation de type *à visée ontologique* dans les SE (Section 7.2.2). Il apparaît que la majorité des traits sont liés à des informations présentes dans l'amorce.

Traits	Informations capturées	Composants	corrélation r
t_NbToken	nombre tokens : 1	Amorce	-0,236
t_POS_c	contient : Verbe conjugué	Item	-0,219
t_POS_c	contient : Nom pluriel	Amorce	0,210
t_POS_c	contient : Préposition	Item	0,210
t_POS_c	contient : Verbe conjugué	Amorce	0,195
t_POS_c	contient : Nom propre	Amorce	0,176
t_Lexique	marqueurs de relation : *holonymie*	Amorce	0,151
t_Saturation_c	Amorce incomplète	Amorce	0,141
t_Lexique	marqueurs de relation : *metalexicale*	Amorce	-0,126
t_NbToken	nombre tokens : 3	Item	0,099

TABLE C.2 : Ordonnancement des dix traits avec les valeurs absolues de corrélation les plus élevées pour le type sémantique *à visée ontologique*

C.5 Stop-liste d'entités textuelles pour l'identification des arguments de la relation

Dans le cadre de l'exploitation des pages Wikipédia, cette stop-liste est utilisée pour filtrer les entités textuelles qui apparaissent isolées dans les amorces (Section 7.3.2). Ces entités textuelles sont celles qui présentent au moins deux occurrences dans notre corpus annoté. Il s'agit essentiellement de titres génériques[1] de Wikipédia ou de circonstants.

- Algérie
- Allemagne
- Articles connexes
- Belgique
- Bibliographie
- Espagne
- États-Unis
- Europe
- Filmographie
- France
- Fruits
- Généralités
- Italie
- Liens externes
- Québec
- Royaume-Uni
- Sciences
- Suisse
- Technique
- Toponyme

[1] https://fr.wikipedia.org/wiki/Aide:Plans_d'articles

Annexe D

Planches de structures énumératives

Sommaire

Ces planches présentent des exemples de structures énumératives verticales rencontrées dans notre corpus annoté (Chapitre 6), ou dans les données d'évaluation pour l'ensemble du système (Section 7.4).

D.1 SE_port

Exemple de structure énumérative porteuse d'une relation d'hyperonymie. Notons que le dernier item rompt le parallélisme et contient une énumération horizontale.

(4.a)

Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets…) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants. Les matériaux extraits sont transportés par une marie-salope.
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.
- Les bateaux de ravitaillement : on trouve notamment les pétroliers ravitailleurs afin de remplir les soutes, et différentes barges pour l'avitaillement lorsque celui-ci n'est pas fait depuis la terre. Les allèges servent à transporter les marchandises entre le quai et le navire, mais ne sont plus guère employées.
- Divers bateaux utilisés pour la sécurité : bateaux-pompe en cas d'incendie, canots de sauvetage pour le secours en mer, patrouilleurs, navires des gardes-côte et navires de l'autorité portuaire

D.2 SE_intertidaux

Exemple de structure énumérative à deux temps (Porhiel, 2007).

(4.b)

Les milieux intertidaux sont des écotones particuliers, dont slikke et schorre sont les deux principales composantes en zone continentale, remplacées par la mangrove en zone tropicale.

- La slikke est l'étage le plus bas : exposée à la mer, zone vaseuse immergée à chaque marée, apparemment pauvre, elle abrite une vie intense, essentiellement des macroinvertébrés et micro-organismes. La basse-slikke, gorgée d'eau, accueille des plantes phanérogames rare (réduite aux zostères). La haute-slikke est, elle, couverte de salicornes et de spartines (graminées dures résistantes au sel).
- Le schorre n'est submergé qu'aux grandes marées et lors des tempêtes, mais il est exposé aux embruns. Il abrite des graminées constituant les prés salés et une végétation d'autant plus variée que l'eau douce est présente.
- Le bas-schorre est un milieu de transition accueillant encore des espèces de la haute slikke qui se mélangent à la glycérie maritime (Puccinellia maritima) et à l' aster maritime.
- Le moyen-schorre accueille l'obione faux-pourpier (sous-arbrisseau aux feuilles persistantes) évoluant vers le haut schorre enrichi de statice maritime (lavande de mer), plantain maritime, avec encore l'aster et la glycérie maritime. Coléoptères, diptères, collemboles complètent la faune des crustacés des bords de slikke, qui nourrissent de nombreux oiseaux (laridés (mouettes et goélands), limicoles, oies bernaches, canards, hérons à marée basse et oiseaux plongeurs piscivores (grèbes) ou malacophages (eiders, macreuses) à marée haute

D.3 SE_digues

Exemple de structure énumérative porteuse d'une relation d'hyperonymie, où l'hyperonyme est contenu dans un titre.

(4.c)

Grands types de digues

On peut distinguer :

- **les digues de protection contre les inondations**. Elles sont situées dans le lit majeur d'un cours d'eau ou le long du littoral, parallèlement à la rive et destinées à contenir les eaux de celui-ci à l'extérieur des digues. Elles portent alors parfois le nom de levée ; c'est ce qu'on trouve, par exemple, sur le Mississippi.
- **les digues de canaux** (d'irrigation, hydroélectriques...), les canaux sont généralement alimentés artificiellement, les digues de canaux servent à contenir l'eau à l'intérieur du canal. Les remblais composant des barrages sont parfois appelés digues (exemple : digue d'étang), mais pour éviter toute confusion, il n'est pas recommandé d'employer le mot digue pour désigner un ouvrage transversal qui barre un cours d'eau ;
- **les jetées** ou digues portuaires, plus ou moins longues faisant à la fois office de brise-lame et d'écran aux vagues. N'ayant qu'une fonction de protection contre les vagues et courants, elles n'ont pas vocation à être étanches ; Certaines digues sont basses et constituées de blocs de pierre qui atténuent les vagues sans empêcher l'eau d'y circuler.
- les **ouvrages de protection contre la mer**, de plus en plus nombreux, et qui constituent par exemple une grande partie du littoral des Pays-Bas, isolant et protégeant les polders de la mer ;

D.4 SE_gaz

Exemple de structure énumérative porteuse d'une relation d'hyperonymie. Notons que la présence de circonstants temporels en début de chaque item, ainsi que la coordination syntaxique dans le dernier item.

(4.d)	Seront également mis en œuvre les gaz suivants : • À partir de 1895, le pétrole vaporisé n'est toutefois pas un gaz mais une vaporisation. • À partir de 1900, l'acétylène est utilisé en France jusqu'aux alentours de 1940. Ce gaz était alors assez dangereux et son stockage demandait l'utilisation de citernes garnies d'un ciment poreux. L'incandescence à l'acétylène sera essayée au phare de Chassiron à titre expérimental de 1902 à 1905. Il sera très utilisé à l'étranger. • En 1923, le gaz BBT (du fabricant français de phares Barbier Bénard Turenne) représente une forte amélioration des qualités de compression et de sécurité. Il sera produit entre les deux guerres mondiales. Des usines de fabrication seront installées à Sfax en Tunisie et à Marseille. • Après 1935, le butane et le propane : Le développement des exploitations pétrolières, dans les années 1930, permettait la fabrication standardisée du propane puis du butane. Les premiers essais en mer seront réalisés au banc du Turc (Phare de la Banche), en face de Lorient en 1932. Il faudra toutefois attendre la fin de la Seconde Guerre mondiale pour qu'une utilisation régulière soit faite par le Service des phares et balises. Les deux gaz seront utilisés jusque dans les années 1980. Les citernes de gaz étaient directement livrées dans les services qui les ventilaient en fonction de leurs besoins.

D.5 SE_blockhaus

Exemple de structure énumérative avec définition d'un terme.

(4.e)	En français courant, blockhaus est devenu un terme générique comme bunker ou casemate et désigne désormais tout type d'ouvrage militaire bétonné, a priori isolé ou de petite dimension. Son équivalent strict est tout simplement bloc, employé pour la Ligne Maginot. Les militaires du génie écrivent aussi bloc bétonné. Ils réservent le terme de blockhaus à : • En fortification de campagne, un retranchement protégé par des rondins et recouvert de terre, abritant des fantassins avec leur armement. • En fortification permanente en pierre (système Séré de Rivières), le réduit d'une position d'infanterie, souvent situé en montagne, parfois en bord de mer, peu susceptible d'être battu par l'artillerie adverse, et constitué soit d'une tour munie de meurtrières, de bretèches ou de mâchicoulis, soit d'un casernement pourvu de volets métalliques percés de meurtrières. • En fortification moderne (type Maginot), soit un petit ouvrage extérieur en béton équipé d'armes automatiques légères — les ouvrages de taille intermédiaire sont nommés casemates et les plus importants blocs — , soit un local intérieur pourvu d'un fusil mitrailleur et placé à un coude d'un couloir de communication afin de le battre en cas d'intrusion de l'adversaire.

D.6 SE_capteur

Exemple d'une structure énumérative porteuse d'une relation d'hyperonymie et où les hyponymes sont précédés d'une clause.

(4.f)

Le guidage à la place est un concept qui permet de trouver immédiatement la place libre de son choix dans un parc de stationnement, même en cas de forte affluence. Le système indique aux automobilistes les places disponibles par zones, par niveaux, et dans les allées de circulation, et apporte à l'exploitant des statistiques très détaillées sur l'occupation du parc. Chaque place de stationnement, est équipée d'un capteur qui détecte la présence des véhicules stationnés et la transmet en temps réel au système. Deux systèmes existent actuellement :

- Le plus fiable et efficace : Des capteurs à ultrasons placés en hauteur au dessus des places avec un voyant lumineux à diodes LED devant chaque place, qui indique aux usagers, en temps réel, les places disponibles (voyant vert) et occupées (voyant rouge). Cette technologie reste aujourd'hui de loin la plus fiable dans les parking couverts. Pour que ce système soit efficace pour les usagers, les voyants à LED doivent être à haute luminosité et omnidirectionnels (visibles sur 360°) et pour être bien visibles dans tout le parking. Le montage en hauteur évite aussi tout risque de chocs ou de vandalisme sur les équipements.
- Dans des cas très particuliers : Des capteurs à induction magnétique placés au sol, qui transmettent l'information par radio-fréquence formant un réseau de capteurs. La technologie RFID sur laquelle repose ce système sans-fil permet un comptage dans les parkings à l'extérieur. Cette solution évite une partie des câblages, mais ne permet pas d'indiquer aux usagers les places libres dans les allées : sa fonction est donc d'indiquer le nombre total des places libres par zone et par allée sur des afficheurs. Mais l'installation de ces afficheurs de comptage en extérieur est complexe et nécessite des travaux de structure importants et onéreux. Enfin le principe de détection de la variations de champ magnétique a une fiabilité limitée car il existe dans les parkings diverses sources de variations de champ magnétique qui créent des perturbations qui génèrent le plus souvent des erreurs de comptage. Ces capteurs au sol fonctionnent sur piles qu'il faut aussi changer après quelques années.

D.7 SE_volcan

Exemple d'une structure énumérative porteuse d'une relation d'hyperonymie.

(4.g)

La classification la plus courante dans les ouvrages de vulgarisation distingue trois types de volcans suivant le type de lave qu'ils émettent et le type d'éruption :

- en volcan bouclier lorsque son diamètre est très supérieur à sa hauteur en raison de la fluidité des laves qui peuvent parcourir des kilomètres avant de s'arrêter ; le Mauna Kea, l'Erta Ale ou le Piton de la Fournaise en sont des exemples ;
- en stratovolcan lorsque son diamètre est plus équilibré par rapport à sa hauteur en raison de la plus grande viscosité des laves ; il s'agit des volcans aux éruptions explosives comme le Vésuve, le mont Fuji, le Merapi ou le mont Saint Helens ;
- en volcan fissural formé par une ouverture linéaire dans la croûte terrestre ou océanique par laquelle s'échappe de la lave fluide ; les volcans des dorsales se présentent sous forme de fissure comme le Laki ou le Krafla.

D.8 SE_atout

Deux structures énumératives imbriquées.

(4.h)

Atouts liés aux volcans

Par certains aspects, l'homme peut tirer profit de la présence des volcans avec :

- l'exploitation de l'énergie géothermique pour production d'électricité, le chauffage des bâtiments ou des serres pour les cultures ;
- la fourniture de matériaux de construction, ou à usage industriel tels que :
  - le basalte qui sert de pierres de construction, de ballast ou de gravas concassé ;
  - la ponce et la pouzzolane qui servent, entre autres, d'isolant dans les bétons ;
  - l'extraction des minerais de soufre, de cuivre, de fer, de platine, de diamants, etc.
- la fertilisation des sols tels les versants de l'Etna qui constituent une région à très forte densité agricole en raison de la fertilité des sols volcaniques et où d'immenses vergers d'agrumes sont implantés. Ces sols volcaniques fertiles font vivre 350 millions de personnes dans le monde.

D.9 SE_transmission

Exemple d'une structure énumérative porteuse d'une relation d'hyperonymie et où les unités logiques directement subordonnées à l'unité logique de l'amorce contiennent des éléments contextuels. Les entités textuelles soulignées sont celles retournées par le système.

(4.i)

- transmission sans fil
  - Courte distance
    - Bluetooth
  - Moyenne distance
    - Wi-Fi, 802.11
    - MANET
  - Longue distance
    - MMDS
    - SMDS
    - Transmission de données sur téléphone cellulaire
      - CDMA
      - CDPD
      - GSM
      - GPRS
      - TDMA
    - Réseaux de téléavertissement
      - DataTAC
      - Mobitex
      - Motient

D.10 SE_transporteur

Exemple d'une structure énumérative porteuse d'une relation d'hyperonymie. Les entités textuelles soulignées sont celles retournées par le système.

(4.j)

Principaux transporteurs frigorifiques

- STEF
- STG (transporteur)
- Groupe Delanchy
- Norbert Dentressangle
- Le Calvez
- Madrias
- Gringore
- Antoine distribution
- Express marée
- Groupe Malherbe
- Groupe Olano

D.11 SE_marchandises

Exemple d'une structure énumérative avec des entités textuelles coordonnées. Les entités textuelles soulignées sont celles retournées par le système.

(4.k)	Transports de marchandises : - Légères : camionnette, fourgonnette, triporteur - Lourdes : poids lourd, semi-remorque, wagon, train, cargo, porte-conteneurs, pétrolier

D.12 SE_sql

Les entités textuelles soulignées sont celles retournées par le système.

(4.l) On distingue typiquement quatre types de commandes :
- CREATE : création de la structure
- ALTER : modification de la structure
- DROP : suppression des données et de la structure
- RENAME : renommage,

D.13 SE_filiales

Les entités textuelles soulignées sont celles retournées par le système.

(4.m)	• Canadien National (CN) et filiales : – BC Rail (BCR) – Duluth, Winnipeg and Pacific Railway (DWP) – Elgin, Joliet and Eastern Railway (EJE) – Grand Trunk Western Railroad (GTW) – Great Lakes Transportation (GLT) – Illinois Central Railroad (IC) – Lakeland & Waterways Railway (LWR) – Mackenzie Northern Railway (MKNR) – Savage Alberta Railnet (SAR) – Wisconsin Central Ltd. (WC)

D.14 SE_compression

Les entités textuelles soulignées sont celles retournées par le système.

(4.n)

On peut distinguer deux types de compression :

- les compressions sans a priori sur les données : ce sont des algorithmes qui travaillent uniquement sur les nombres, quelle que soit l'information portée par ces nombres ; ils sont donc généraux, pas spécifiques aux données ; on peut distinguer :
  - les algorithmes à table stockée : l'algorithme fait une première analyse pour repérer les éléments se répétant, et construit une table de correspondance avec un code raccourci pour chaque élément répétitif ; la taille occupée par la table de stockage fait que ce procédé est plutôt adapté aux gros fichiers,
  - les algorithmes à table construite à la volée : la table de correspondance est construite de manière systématique, sans analyse préalable du fichier ; elle peut être reconstruite à la volée à partir du fichier compressé ; c'est par exemple le cas de l'algorithme de Lempel-Ziv-Welch (LZW) ;
- les compressions spécifiques aux données : si l'on connaît les données, on peut optimiser l'algorithme ; par exemple si l'on sait que l'on a affaire à un texte, on peut se baser sur la fréquence d'utilisation des mots dans le langage ; on distingue deux sous-catégories :
  - les compressions sans perte d'informations,
  - les compressions avec perte de données : la première idée est de faire un « sous-échantillonnage », c'est-à-dire de simplement dégrader la qualité des données en étudiant les sens et la manière dont le cerveau interprète les informations, on peut dégrader certaines caractéristiques des données peu sensibles, donc sans trop altérer la qualité globale des données ; ainsi, si l'oreille humaine est peu sensible à certaines gammes de fréquences, on peut dégrader (voire supprimer) certaines parties du spectre et pas d'autres (MP3) ;
    les algorithmes de compression d'image (JPEG) et de film (MPEG) utilisent une perte de qualité.

Bibliographie

Abu-Mostafa, Y. S., Magdon-Ismail, M. et Lin, H.-T. (2012). *Learning from data*. AMLBook.

Adam, C. (2012). *Voisinage lexical pour l'analyse du discours*. Thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Adam, J.-M. (2011). *Les textes : types et prototypes : récit, description, argumentation, explication et dialogue*. Armand Colin.

Adam, J.-M. et Revaz, F. (1989). Aspects de la structuration du texte descriptif : les marqueurs d'énumération et de reformulation. *Langue française*, pages 59–98.

Adobe (1985). *PostScript language reference manual*. Addison-Wesley Longman Publishing Co., Inc.

Adobe (1992a). Encapsulated postscript file format specification. Rapport technique.

Adobe (1992b). Postscript language document structuring conventions specification. Rapport technique.

Adobe (2000). Pdf reference - second edition. Rapport technique.

Adobe (2001). Pdf reference - third edition. Rapport technique.

Adobe (2008). Document management – portable document format : Pdf 1.7. Rapport technique.

Afantenos, S., Denis, P., Muller, P. et Danlos, L. (2010). Learning recursive segments for discourse parsing. *In Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*.

Afantenos, S. D., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., J., R., T., T., M., V.-C. et L., V. (2012). An empirical resource for discovering cognitive principles of discourse organisation : the annodis corpus. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 2727–2734.

Agichtein, E. et Gravano, L. (2000). Snowball : Extracting relations from large plain-text collections. *In Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Aizerman, A., Braverman, E. M. et Rozoner, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837.

Akbik, A., Visengeriyeva, L., Herger, P., Hemsen, H., Löser, A. *et al.* (2012). Unsupervised discovery of relations and discriminative extraction patterns. *In COLING*, pages 17–32.

Albert, A. et Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Alfonseca, E., Castells, P., Okumura, M. et Ruiz-Casado, M. (2006). A rote extractor with edit distance-based generalisation and multi-corpora precision calculation. *In Proceedings of the COLING/ACL on Main conference poster sessions*, pages 9–16. Association for Computational Linguistics.

Allen, P., Bateman, J. A. et Delin, J. (1999). Genre and layout in multimodal documents : towards an empirical account. *In Power, R. and Scott, D., editors, Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding, or Retrieval of Docu-38 Generating Text, Diagrams and Layout Appropriately According to Genre—John A. Bateman and Renate Henschel.*

André, J., Furuta, R. K. et Quint, V. (1989). *Structured documents*, volume 2. Cambridge University Press.

André, J., Quint, V. *et al.* (1990). Structures et modèles de documents. *Le document électronique.*

Arnold, P. et Rahm, E. (2014). Extracting semantic concept relations from wikipedia. *In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, page 26. ACM.

Artstein, R. et Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asher, N. (1993). *Reference to Abstract Objects in Discourse.* Kluwer Academic Publishers, Dordrecht.

Asher, N. et Lascarides, A. (2003). *Logics of conversation.* Cambridge University Press.

Asher, N. et Vieu, L. (2005). Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610.

Aubin, S. et Hamon, T. (2006). Improving term extraction with terminological resources. *In Advances in Natural Language Processing*, pages 380–387. Springer.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. et Ives, Z. (2007). Dbpedia : A nucleus for a web of open data. *In The semantic web*, pages 722–735. Springer.

Auer, S. et Lehmann, J. (2007). What have innsbruck and leipzig in common ? extracting semantics from wiki content. *In The Semantic Web : Research and Applications*, pages 503–517. Springer.

Auger, A. et Barrière, C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, 14(1):1–19.

Aussenac-Gilles, N. et Kamel, M. (2009). Ontology learning by analyzing xml document structure and content. *In KEOD*, pages 159–165.

Aussenac-Gilles, N. et Séguéla, P. (2000). Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*, (25):175–198.

Austin, J. L. (1975). *How to do things with words*. The William James Lectures. Harvard University Press, second edition édition.

Aït-Mokhtar, S., Lux, V. et Bánik, E. (2003). Linguistic parsing of lists in structured documents. *In Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003), Budapest, Hungary*, volume 9. Citeseer.

Bach, N. et Badaskar, S. (2007). A survey on relation extraction.

Baker, C., Fillmore, C. et Lowe, J. (1998). The berkeley framenet project. *In Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, page 86–90, Montréal.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M. et Etzioni, O. (2007). Open information extraction for the web. *In IJCAI*, volume 7, pages 2670–2676.

Bar-Hillel, Y., Gaifman, C. et Shamir, E. (1960). On categorial and phrase structure grammars. *Bulletin of the research council of Israel*, 9.

Baroni, M., Bernardini, S., Ferraresi, A. et Zanchetta, E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Barque, L., Nasr, A. et Polguere, A. (2010). From the de finitions of the trésor de la langue française to a semantic database of the french language. *In Proceedings of the 14th EURALEX International Congress*.

Bateman, J. et Delin, J. (2001). From genre to text critiquing in multimodal documents. *In Workshop on Multidisciplinary Approaches to Discourse : Improving Text : From text structure to text type*, pages 5–8.

Bateman, J., Kamps, T., Kleinz, J. et Reichenberger, K. (2001). Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, 27(3):409–449.

Bateman, J. et Teich, E. (1995). Selective information presentation in an integrated publication system : an application of genre-driven text generation. *Information processing & management*, 31(5):753–767.

Ben-Hur, A. et Weston, J. (2010). A user's guide to support vector machines. *In Data mining techniques for the life sciences*, pages 223–239. Springer.

Bengio, Y. (2009). Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1):1–127.

Bennett, E. M., Alpert, R. et Goldstein, A. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308.

Berger, A. L., Pietra, V. J. D. et Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Berland, M. et Charniak, E. (1999). Finding parts in very large corpora. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics.

Berners-Lee, T. et Connolly, D. (1993). Hypertext markup language (html) - a representation of textual information and metainformation for retrieval and interchange. Rapport technique, IETF IIIR Working Group.

Berners-Lee, T., Hendler, J., Lassila, O. *et al.* (2001). The semantic web. *Scientific american*, 284(5):28–37.

Biber, D., Connor, U. et Upton, T. A. (2007). *Discourse on the move : Using corpus analysis to describe discourse structure*, volume 28. John Benjamins Publishing.

Bishop, C. (2006). *Pattern recognition and machine learning*, volume 1. Springer.

Bloomfield, L. (1933). *Language*. New York, réd. 1962 édition.

Bordea, G., Buitelaar, P., Faralli, S. et Navigli, R. (2015). Semeval-2015 task 17 : Taxonomy extraction evaluation (texeval). *In Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Boser, B. E., Guyon, I. M. et Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

Bouayad-Agha, N., Power, R. et Scott, D. (2000). Can text structure be incompatible with rhetorical structure ? *In Proceedings of the first international conference on Natural language generation-Volume 14*, pages 194–200. Association for Computational Linguistics.

Bouayad-Agha, N., Scott, D. et Power, R. (2001). The influence of layout on the interpretation of referring expressions. *In Procceedings of Multidisciplinary Approaches to Discourse (MAD)*, pages 133–141. Stichting Neerlandistiek VU.

Boudin, F. (2013). Taln archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue. *In Traitement Automatique des Langues Naturelles (TALN)*, pages 507–514.

Bounhas, I. et Slimani, Y. (2010). A hierarchical approach for semi-structured document indexing and terminology extraction. *In Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on*, pages 315–320. IEEE.

Bourigault, D. (1994). *Lexter : un Logiciel d'EXtraction de TERminologie : application à l'acquisition des connaissances à partir de textes.* Thèse de doctorat, EHESS.

Bras, M., Prévot, L. et Vergez-Couret, M. (2008). Quelles relations de discours pour les structures énumératives ? *In Congrès Mondial de Linguistique Française*, page 179. EDP Sciences.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. et Yergeau, F. (1998). Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210. http ://www. w3. org/TR/1998/REC-xml-19980210*, 16.

Breuel, T. M. (2002). Two geometric algorithms for layout analysis. *In Document analysis systems v*, pages 188–199. Springer.

Brin, S. (1998). Extracting patterns and relations from the world wide web. *In The World Wide Web and Databases, International Workshop WebDB'98*, pages 172–183, Valencia.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

Brunzel, M. (2008). The xtreem methods for ontology learning from web documents. *In* Buitelaar, P. et Cimiano, P., éditeurs : *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, volume 167 de *Frontiers in Artificial Intelligence and Applications*, pages 3–26. IOS Press, Amsterdam.

Brunzel, M. et Spiliopoulou, M. (2006). Discovering multi terms and co-hyponymy from xhtml documents with xtreem. *In* Nayak, R. et Zaki, M. J., éditeurs : *Knowledge discovery from XML documents.* Springer.

Bunescu, R. C. et Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. *In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731. Association for Computational Linguistics.

Bush, C. (2003). Des déclencheurs des énumérations d'entités nommées sur le web. *Revue québécoise de linguistique*, 32(2):47–81.

Béchet, N., Cellier, P., Charnois, T., Crémilleux, B. et Quiniou, S. (2013). Sdmc : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. *In Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*, Toulouse.

Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E. et Zhang, Y. (2008). Webtables : exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1):538–549.

Candito, M., Crabbé, B., Denis, P. et Guérin, F. (2009). Analyse syntaxique du français : des constituants aux dépendances. *In Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*.

Cartier, E. (2015). Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français. *In Actes de la 22e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, pages 131–145, Caen. Association pour le Traitement Automatique des Langues.

Cattoni, R., Coianiz, T., Messelodi, S. et Modena, C. M. (1998). Geometric layout analysis techniques for document image understanding : a review. *ITC-irst Technical Report*, 9703(09).

Cawley, G. C. et Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107.

Chang, C.-C. et Lin, C.-J. (2013). Libsvm : a library for support vector machines. Rapport technique, Department of Computer Science. National Taiwan University, Taipei, Taiwan.

Charolles, M. (1997). L'encadrement du discours. *In Cahier de Recherche Linguistique*, volume 6. Université de Nancy.

Chernov, S., Iofciu, T., Nejdl, W. et Zhou, X. (2006). Extracting semantics relationships between wikipedia categories. *SemWiki*, 206.

Chinchor, N. et Marsh, E. (1998). Muc-7 information extraction task definition. *In Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pages 359–367.

Chinchor, N. et Robinson, P. (1997). Muc-7 named entity task definition. *In Proceedings of the 7th Conference on Message Understanding*, page 29.

Choi, F. Y. Y. (2002). *Content-based Text Navigation*. Thèse de doctorat, the University of Manchester.

Chomsky (1957). *Syntactic Structures*. Mouton Publisher.

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3):113–124.

Cimiano, P., Handschuh, S. et Staab, S. (2004). Towards the self-annotating web. *In Proceedings of the 13th international conference on World Wide Web*, pages 462–471. ACM.

Cimiano, P., Hotho, A. et Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, 24:305–339.

Clark, J. (1997). Comparison of sgml and xml : World wide web consortium note. Rapport technique, World Wide Web Consortium.

Cohen, J. *et al.* (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Collins, M. (2002). Discriminative training methods for hidden markov models : Theory and experiments with perceptron algorithms. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

Collins, M. et Duffy, N. (2001). Convolution kernels for natural language. *In Advances in neural information processing systems*, pages 625–632.

Condamines, A. (2003). Vers la définition de genres interprétatifs. *Actes de TIA'2003*, pages 69–79.

Condamines, A. (2008). Taking genre into account when analysing conceptual relation patterns. *Corpora*, 3(2):115–140.

Condamines, A. et Rebeyrolle, J. (1997). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. *In Journées Ingénierie des Connaissances et Apprentissage Automatique (JICAA'97)*, pages 191–206.

Constantin, A., Pettifer, S. et Voronkov, A. (2013). Pdfx : fully-automated pdf-to-xml conversion of scientific literature. *In Proceedings of the 2013 ACM symposium on Document engineering*, pages 177–180. ACM.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3): 273–297.

Couto, J., Ferret, O., Grau, B., Hernandez, N., Jackiewicz, A., Minel, J.-L. et Porhiel, S. (2004). Régal, un système pour la visualisation sélective de documents. *Revue d'intelligence artificielle*, 18(4):481–514.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 411–421.

Cox, D. R. et Snell, E. J. (1989). *Analysis of binary data*, volume 32. CRC Press.

Craven, M., Kumlien, J. *et al.* (1999). Constructing biological knowledge bases by extracting information from text sources. *In ISMB*, volume 1999, pages 77–86.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Cruse, D. A. (2002). *Hyponymy and Its Varieties*, chapitre 1, pages 3 – 22. Kluwer Academic Publishers.

Culotta, A. et Sorensen, J. (2004). Dependency tree kernels for relation extraction. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

Daelemans, W. et Hoste, V. (2002). Evaluation of machine learning methods for natural language processing tasks. *In 3rd International conference on Language Resources and Evaluation (LREC 2002)*. European Language Resources Association (ELRA).

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act : Combining symbolic and statistical approaches to language*, 1:49–66.

Damamme-Gilbert, B. (1989). *La série énumérative : étude linguistique et stylistique s' appuyant sur dix romans français publiés entre 1945 et 1975*, volume 19. Librairie Droz.

Darroch, J. et Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, 43(5):1470–1480.

Daumé III, H. (2006). *Practical Structured Learning Techniques for Natural Language Processing*. Thèse de doctorat, University of Southern California, Los Angeles, CA.

De Saussure, F. (1995). *Cours de linguistique générale*. Payot, (1916) édition.

Delin, J., Bateman, J. et Allen, P. (2002). A model of genre in document layout. *Information Design Journal*, 11(1):54–66.

Desclés, J.-P. (1990). *Langages applicatifs, langues naturelles et cognition*. Hermès.

Desclés, J.-P. (2006). Contextual exploration processing for discourse and automatic annotations of texts. *In FLAIRS Conference*, volume 281, page 284.

Dikovsky, A. et Modina, L. (2000). Dependencies on the other side of the curtain. *Traitement Automatique des Langues (TAL)*, 41(1):79–111.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. et Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. *In LREC*.

Dohrn, H. et Riehle, D. (2011). Design and implementation of the sweble wikitext parser : unlocking the structured data of wikipedia. *In Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 72–81. ACM.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S. et Zhang, W. (2014). Knowledge vault : A web-scale approach to probabilistic knowledge fusion. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM.

Dutrey, C., Clavel, C., Rosset, S., Vasilescu, I. et Adda-Decker, M. (2012). Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ? *In Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*.

Ehrmann, M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Paris 7.

Esposito, F., Malerba, D. et Semeraro, G. (1994). Multistrategy learning for document recognition. *Applied Artificial Intelligence an International Journal*, 8(1):33–84.

Estabrooks, A., Jo, T. et Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. et Yates, A. (2004). Web-scale information extraction in knowitall. *In Proceedings of the 13th international conference on World Wide Web*, pages 100–110. ACM.

Faber, P. et L'Homme, M.-C. (2014). Lexical semantic approaches to terminology. an introduction. *Terminology : international journal of theoretical and applied issues in specialized communication*, 20(2):143–150.

Faessel, N. (2011). *Indexation et interrogation de pages Web décomposées en blocs visuels*. Thèse de doctorat, Université Paul Cézanne Aix-Marseille III.

Falco, M.-H. (2014). *Répondre à des questions à réponses multiples sur le Web*. Thèse de doctorat, Université Paris-Sud.

Fauconnier, J.-P. et Kamel, M. (2015). Discovering hypernymy relations using text layout. *In Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 249–258, Denver, Colorado. Association for Computational Linguistics.

Fauconnier, J.-P., Kamel, M. et Rothenburger, B. (2013a). Une typologie multidimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques. *In International Conference on Terminology and Artificial Intelligence (TIA 2013)*.

Fauconnier, J.-P., Kamel, M. et Rothenburger, B. (2015). A Supervised Machine Learning Approach for Taxonomic Relation Recognition through Non-linear Enumerative Structures (papier court). *In ACM Symposium on Applied Computing (SAC 2015)*, Salamanque.

Fauconnier, J.-P., Kamel, M., Rothenburger, B. et Aussenac-Gilles, N. (2013b). Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles. *In Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, pages 132–145.

Fauconnier, J.-P., Sorin, L., Kamel, M., Mojahid, M. et Aussenac-Gilles, N. (2014). Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux. *In Actes de la 21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, pages 340–351.

Faure, D. et Nedellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : The system asium. *In Knowledge Acquisition, Modeling and Management*, pages 329–334. Springer.

Feigenbaum, E. A. et McCorduck, P. (1983). *The fifth generation*. Addison-Wesley Pub.

Fellbaum, C. (1998a). Introduction. *In* Fellbaum, C., éditeur : *WordNet : An Electronic Lexical Database*. Wiley Online Library.

Fellbaum, C. (1998b). *WordNet : An Electronic Lexical Database*. Wiley Online Library.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G. et Jacquemin, C. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *In Actes de la 8e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*.

Fillmore, C. (1982). Frame semantics. *Linguistics in the morning calm*, pages 111–137.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2):139–172.

Fleiss, J. L., Nee, J. C. et Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5):974–977.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.

Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard University Press.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and brain sciences*, 3(01):63–73.

Forgy, E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21:768–769.

Furuta, R. (1989). Concepts and models for structured documents. *In* André, J., Furuta, R. et Quint, V., éditeurs : *Structured Documents*, pages 161–180. Cambridge Series on Electronic Publishing.

Furuta, R., Scofield, J. et Shaw, A. (1982). Document formatting systems : survey, concepts, and issues. *ACM Computing Surveys (CSUR)*, 14(3):417–472.

Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and control*, 8(3):304–337.

Gala, N. (2003). *Un modèle d'analyseur synthaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires.* Thèse de doctorat, Paris 11, Orsay.

Ganapathi, V., Vickrey, D., Duchi, J. et Koller, D. (2008). Constrained approximate maximum entropy learning of markov random fields. *In Conference on uncertainty in artificial intelligence (UAI)*.

Gangemi, A., Guarino, N. et Oltramari, A. (2001). Conceptual analysis of lexical taxonomies : The case of wordnet top-level. *In Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 285–296. ACM.

Garcia, D. (1997). Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes. *Actes des deuxièmes rencontres—Terminologie et Intelligence Artificielle, TIA'97*, pages 7–26.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26.

Goldfarb, D. et Idnani, A. (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. *In Numerical Analysis*, pages 226–239. Springer.

Grabar, N. et Hamon, T. (2004). Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'intelligence artificielle*, 18(1):57–85.

Grabar, N., Malaisé, V., Marcus, A. et Krul, A. (2004). Repérage de relations terminologiques transversales en corpus. *In Actes de la 11ème conférence sur le Traitement Automatique des Langues Naturelles*, Fès, Maroc. Association pour le Traitement Automatique des Langues.

Grabar, N. et Zweigenbaum, P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. *Actes de TALN*, 1999:175–184.

Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. *In Proceedings of the 22nd annual SAS User Group International conference*, pages 1110–1115.

Green, R., Bean, C. A. et Myaeng, S. H. (2002). *The semantics of relationships : an interdisciplinary perspective*, volume 3. Springer.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Springer Science & Business Media.

Grefenstette, G. (2015). Inriasac : Simple hypernym extraction methods. *In SemEval*.

Grishman, R. et Sundheim, B. (1996). Message understanding conference-6 : A brief history. *In COLING*, volume 96, pages 466–471.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O. et Quintard, L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. *In Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100. Association for Computational Linguistics.

GuoDong, Z., Jian, S., Jie, Z. et Min, Z. (2005). Exploring various knowledge in relation extraction. *In Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics.

Ha, J., Haralick, R. M. et Phillips, I. T. (1995a). Document page decomposition by the bounding-box projection technique. *In International Conference Document Analysis and Recognition (ICDAR)*, volume 2, pages 1119–1122.

Ha, J., Haralick, R. M. et Phillips, I. T. (1995b). Recursive xy cut using bounding boxes of connected components. *In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 2, pages 952–955. IEEE.

Halliday, M. A. K. (1977). Text as semantic choice in social contexts. *Grammars and descriptions*.

Halliday, M. A. K. et Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hamon, T. et Nazarenko, A. (2001). Exploitation de l'expertise humaine dans un processus de constitution de terminologie. *In Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 213–222, Tours, France. Association pour le Traitement Automatique des Langues.

Hardy, M. R. et Brailsford, D. F. (2002). Mapping and displaying structural transformations between xml and pdf. *In Proceedings of the 2002 ACM symposium on Document engineering*, pages 95–102. ACM.

Harris, Z. (1968). *Mathematical structures of language*. Numéro 21. John Wiley & Sons, Inc.

Harris, Z. S. (1961). *Structural Linguistcs*. THe University of Chicago Press, Chicago.

Harris, Z. S. (2002). The background of transformational and metalanguage analysis. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 1–18.

Hasegawa, T., Sekine, S. et Grishman, R. (2004). Discovering relations among named entities from large corpora. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 415. Association for Computational Linguistics.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, volume 2. Springer.

Hays, D. G. (1964). Dependency theory : A formalism and some observations. *Language*, pages 511–525.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545. Association for Computational Linguistics.

Hearst, M. A. (1998). Automated discovery of wordnet relations. *In* Fellbaum, C., éditeur : *WordNet : An Electronic Lexical Database*. Wiley Online Library.

Hellwig, P. (2006). Parsing with dependency grammar. *In* Ágel V., éditeur : *Dependency and valency : an international handbook of contemporary research*, volume 2, chapitre 79. Walter de Gruyter.

Hernandez, N. et Grau, B. (2005). Détection automatique de structures fines de texte. *In Actes de la 12e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*.

Hickson, I., Berjon, R., Faulkner, S., Leithead, T., Navara, E. D., O'Connor, E. et Pfeiffer, S. (2014). A vocabulary and associated apis for html and xhtm. Rapport technique, World Wide Web Consortium, Working Draft.

Hirst, G. (2009). Ontology and the lexicon. *In Handbook on ontologies*, pages 269–292. Springer.

Ho-Dac, L.-M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. Thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P. et Rebeyrolle, J. (2009). Des indices aux marqueurs : méthodes de découverte de marqueurs discursifs complexes. *In Linguistic and Psycholinguistic Approaches to Text Structuring*.

Ho-Dac, L.-M., Jacques, M.-P. et Rebeyrolle, J. (2004). Sur la fonction discursive des titres. *L'unité texte, Pleyben, Perspectives*, pages 125–152.

Ho-Dac, L.-M., Péry-Woodley, M.-P. et Tanguy, L. (2010). Anatomie des structures énumératives. *In Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*.

Hoffart, J., Suchanek, F. M., Berberich, K. et Weikum, G. (2013). Yago2 : A spatially and temporally enhanced knowledge base from wikipedia. *In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3161–3165. AAAI Press.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. et Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Honeste, M.-L. et Froissart, C. (2003). Blancs, casses, puces, tirets,.... *Ordre et distinction dans la langue et le discours*, pages 252–276.

Hovy, E. H. et Arens, Y. (1991). Automatic generation of formatted text. *In Proceedings of the 9th AAAI Conference (AAAI 1991)*, Anaheim, CA.

International Organization for Standardization (1987). *ISO 704 : Principles and methods of terminology*.

International Organization for Standardization (1989). *ISO 8613 : Information Processing - Text and office systems - Office Docment Architecture (ODA) and Interchange Format*.

International Organization for Standardization (1990). *ISO 1087 : Terminology – Vocabulary*.

Jackendoff, R. (1990). *Semantic structures*, volume 18 de *Current Studies in Linguistics*. MIT press.

Jackiewicz, A. (2005). Les séries linéaires dans le discours. *Langue Française*, 148:95–110.

Jackiewicz, A. et Minel, J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. *TALN 2003*, pages 95–107.

Jacquemin, C. (1994). Fastr : A unification-based front-end to automatic indexing. *In RIAO 94 : recherche d'information assistée par ordinateur. Conférence*, pages 34–47.

Jacquemin, C. (1996). A symbolic and surgical acquisition of terms through variation. *In Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438. Springer.

Jacquemin, C. (1997). Guessing morphology from terms and corpora. *In ACM SIGIR Forum*, volume 31, pages 156–165. ACM.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*. MIT press.

Jacquemin, C. et Bush, C. (2000). Fouille du web pour la collecte d'entités nommées. *Actes de TALN*, pages 187–196.

Jacques, M.-P. et Aussenac-Gilles, N. (2006). Variabilité des performances des outils de tal et genre textuel. *Traitement automatique des langues*, 47(1):11–32.

Jain, A. K. et Bhattacharjee, S. (1992). Text segmentation using gabor filters for automatic document processing. *Machine Vision and Applications*, 5(3):169–184.

Jannink, J. (1999). Thesaurus entry extraction from an on-line dictionary. *In Proceedings of Fusion*, volume 99. Citeseer.

Jansche, M. (2005). Maximum expected f-measure training of logistic regression models. *In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 692–699. Association for Computational Linguistics.

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical review*, 106(4): 620.

Joachims, T. (2002). *Learning to classify text using support vector machines : Methods, theory and algorithms*. Kluwer Academic Publishers.

Jouis, C. (2002). Logic of relationships. *In* Green, R., Bean, C. A. et Myaeng, S. H., éditeurs : *The semantics of relationships : an interdisciplinary perspective*, pages 127–140. Springer.

Jouis, C., Biskri, I., Descles, J.-P., Le Priol, F., Meunier, J.-G., Mustafa, W. et Nault, G. (1997). Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques. *In Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, Avignon.

Jousse, F., Gilleron, R., Tellier, I. et Tommasi, M. (2006). Conditional random fields for xml trees. *In Workshop on Mining and Learning in Graphs*.

Kahane, S. (2000). Présentation. *TAL (Traitement automatique des langues)*, 41(1):7–13.

Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics.

Kamel, M. et Rothenburger, B. (2011). Elicitation de structures hiérarchiques à partir de structures enumératives pour la construction d'ontologie. *In Journées Francophones d'Ingénierie des Connaissances (IC 2011)*, pages 505–522, Annecy.

Kamel, M., Rothenburger, B. et Fauconnier, J.-P. (2014). Identification de relations sémantiques portées par les structures énumératives paradigmatiques. *Revue d'Intelligence Artificielle*, Ingénierie des Connaissances.

Katz, J. J. et Fodor, J. A. (1963). The structure of a semantic theory. *Language*, (39):170–210.

Kergosien, E., Kamel, M., Sallaberry, C., Bessagnet, M.-N., Aussenac-Gilles, N. et Gaio, M. (2010). Construction et enrichissement automatique d'ontologie à partir de ressources externes.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, A. et Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Kim, K., Jung, K. et Kim, H. (2005). Fast color texture-based object detection in images : Application to license plate localization. *In Support Vector Machines : Theory and Applications*, pages 297–320. Springer.

Kise, K., Sato, A. et Iwata, M. (1998). Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382.

Klink, S., Dengel, A. et Kieninger, T. (2000). Document structure analysis based on layout and textual features. *In Proc. of International Workshop on Document Analysis Systems, DAS2000*, pages 99–111. Citeseer.

Knuth, E. D. (1965). On the translation of languages from left to right. *Information and control*, 8(6):607–639.

Knuth, E. D. (1984). *The TeXbook*. Addison-Wesley Reading, Massachusetts.

Knuth, E. D. (1995). The dvitype processor (version 3.6, december 1995). *CTAN*.

Kopec, G. E., Chou, P. *et al.* (1994). Document image decoding using markov source models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6):602–617.

Kotsiantis, S. B., Zaharakis, I. et Pintelas, P. (2007). Supervised machine learning : A review of classification techniques. *Informatica*, 31:249–268.

Kripke, S. A. (1982). *Naming and necessity*. Harvard University Press, Boston.

Krippendorff, K. (1980). *Content analysis : An introduction to its methodology*. Sage Publications.

Krishnamoorthy, M., Nagy, G., Seth, S. et Viswanathan, M. (1993). Syntactic segmentation and labeling of digitized pages from technical journals. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(7):737–747.

Kruschwitz, U. (2001). Exploiting structure for intelligent web search. *In System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on*, pages 9–pp. IEEE.

Kübler, S., McDonald, R. et Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127.

Lafferty, J., McCallum, A. et Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *Department of Computer & Information Science, University of Pennsylvania*.

Laignelet, M. (2009). *Analyse discursive pour le reperage automatique de segments obsolescents dans des documents encyclopediques.* Thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Laignelet, M., Kamel, M. et Aussenac-Gilles, N. (2011). Enrichir la notion de patron par la prise en compte de la structure textuelle - application à la construction d'ontologie. *In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France. Association pour le Traitement Automatique des Langues. Enriching the notion of pattern by taking into account the textual structure - Application to ontology construction.

Lambrou-Latreille, K. (2015). Relation extraction pattern ranking using word similarity. *In NAACL-HLT 2015 Student Research Workshop (SRW)*, page 25.

Lamport, L. (1994). *LATEX : A Document Preparation System*. Addison-Wesley, Reading.

Le Priol, F. (2001). Identification, interprétation et représentation de relations sémantiques entre concepts. *In Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 373–378, Tours, France. Association pour le Traitement Automatique des Langues.

Lehrer, A. (1974). Semantic fields and lexical structure.

Lenci, A. (2001). Building an ontology for the lexicon : Semantic types and word meaning. *In Ontology-Based Interpretation of Noun Phrases : Proceedings of the First International OntoQuery Workshop, University of Southern Denmark*, pages 103–120.

Lerat, P. (2009). La combinatoire des termes. exemple : nectar de fruits. *Hermes. Journal of Language and Communication Studies*, pages 211–232.

Levy, R. et Andrew, G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures. *In Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234. Citeseer.

Liu, D. C. et Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Luc, C. (1998). Types de contraintes architecturales sur la composition d'objets textuels. *In Actes, CIDE'98 (Colloque International sur le Document Électronique)*, pages 15–30.

Luc, C. (2000). *Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés.* Thèse de doctorat, Université Paul Sabatier.

Luc, C. (2001). Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. *In Actes de la 8e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, pages 263–272.

Luenberger, D. G. (1984). Linear and nonlinear programming. *Addison-Wesley.*

Lüngen, H., Bärenfänger, M., Hilbert, M., Lobin, H. et Puskás, C. (2010). Discourse relations and document structure. *In Linguistic modeling of information and markup languages*, pages 97–123. Springer.

Lyons, J. (1977). *Semantics.* Cambridge University Press, Cambridge.

Malaisé, V., Zweigenbaum, P. et Bachimont, B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *In Actes de la 11e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004)*, pages 269–278.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. *In proceedings of the 6th conference on Natural language learning*, pages 1–7. Association for Computational Linguistics.

Manabe, T. et Tajima, K. (2015). Extracting logical hierarchical structure of html documents based on headings. *Proceedings of the VLDB Endowment*, 8(12):1606–1617.

Mann, W. C. et Thompson, S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, 8(3):243–281.

Mao, S., Rosenfeld, A. et Kanungo, T. (2003). Document structure analysis algorithms : a literature survey. *In Electronic Imaging 2003*, pages 197–207. International Society for Optics and Photonics.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 365–372. Association for Computational Linguistics.

Mathet, Y. et Widlöcher, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *In Actes de la 18e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*.

Maurel, F., Luc, C., Mojahid, M., Vigouroux, N., Virbel, J. et Nespoulous, J.-L. (2002). Problématique du traitement des structures visuelles dans la présentation oralisée des textes. *In Documents Virtuels Personnalisables 2002 (DVP 2002), Brest, 10/07/2002-11/07/2002*, pages 11–24. ENST Brest.

McCallum, A., Freitag, D. et Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. *In ICML*, pages 591–598.

McDonald, R., Crammer, K. et Pereira, F. (2005a). Online large-margin training of dependency parsers. *In Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 91–98. Association for Computational Linguistics.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y. et White, P. (2005b). Simple algorithms for complex relation extraction with applications to biomedical ie. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 491–498. Association for Computational Linguistics.

Mel'čuk, I. A. (1988). *Dependency syntax : theory and practice*. SUNY Press.

Mel'čuk, I. A. et Wanner, L. (1996). Lexical functions and lexical inheritance for emotion lexemes in german. *In* Wanner, L., éditeur : *Lexical functions in lexicography and natural language processing*, page 209. John Benjamins Publishing, Amsterdam.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2:279.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, G. A. (1990). Nouns in wordnet : a lexical inheritance system. *International journal of Lexicography*, 3(4):245–264.

Miller, G. A. et Hristea, F. (2006). Wordnet nouns : Classes and instances. *Computational linguistics*, 32(1):1–3.

Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R. et Weischedel, R. (1998). Algorithms that learn to extract information - bbn : Description of the sift system as used for muc-7. *In Proceedings of the 7th Conference on Message Understanding*, pages 75–89. Association for Computational Linguistics.

Minsky, M. (1975). A framework for representing knowledge. *In* Winston, P. H., éditeur : *The psychology of computer vision*, page 211. McGraw-Hill, New York.

Mintz, M., Bills, S., Snow, R. et Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Morin, E. (1998). Prométhée, un outil d'aide à l'acquisition de relations sémantiques entre termes. *In Actes de la 5e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 1998)*, pages 172–181.

Morin, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Université de Nantes.

Morin, F. et Bengio, Y. (2005). Hierarchical probabilistic neural network language model. *In Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252. Citeseer.

Morris, J. et Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Müller, S. (2015). *Grammatical Theory : From Transformational Grammar to Constraint-Based Approaches*. Numéro 1 de Lecture Notes in Language Sciences. Language Science Press, Berlin. Open Review Version.

Murphy, M. L. (2003). *Semantic relations and the lexicon : Antonymy, synonymy and other paradigms*. Cambridge University Press.

Myers, E. W. (1986). An o(nd) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266.

Müller, C. et Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *In* Braun, S., Kohn, K. et Mukherjee, J., éditeurs : *Corpus Technology and Language Pedagogy : New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., Tzu-Yi, K., Magistry, P. et Chu-Ren, H. (2009). Wiktionary and nlp : Improving synonymy networks. *In Proceedings of the 2009 Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, pages 19–27. Association for Computational Linguistics.

Navigli, R. et Velardi, P. (2007). Glossextractor : A web application to automatically create a domain glossary. *In AI*IA 2007 : Artificial Intelligence and Human-Oriented Computing*, volume 4733 de *Lecture Notes in Computer Science*, pages 339–349. Springer.

Navigli, R. et Velardi, P. (2008). From glossaries to ontologies : Extracting semantic structure from textual definitions. *In* Buitelaar, P. et Cimiano, P., éditeurs : *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, volume 167 de *Frontiers in Artificial Intelligence and Applications*, pages 71–104. IOS Press, Amsterdam.

Nayak, R. et Zaki, M. J. (2006). *Knowledge discovery from XML documents*. Numéro 3915 de Lecture Notes in Computer Science. Springer.

Nazarenko, A. (2005). Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? *Sémantique et corpus*, pages 211–244.

Nazarenko, A. et Hamon, T. (2002). Structuration de terminologie : quels outils pour quelles pratiques ? *TAL. Traitement automatique des langues*, 43(1):7–18.

Ng, A. Y. et Jordan, A. (2002). On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841.

Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Nobata, C., Collier, N. et Tsujii, J. (2000). Comparison between tagged corpora for the named entity task. *In Proceedings of the workshop on Comparing corpora*, pages 20–27. Association for Computational Linguistics.

Nunberg, G. (1990). *The linguistics of punctuation*. Numéro 18. CSLI Publications.

O'Gorman, L. (1993). The document spectrum for page layout analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(11):1162–1173.

Oh, J.-H., Uchimoto, K. et Torisawa, K. (2009). Bilingual co-training for monolingual hyponymy-relation acquisition. *In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNL*, pages 432–440. Association for Computational Linguistics.

Oh, J.-H., Yamada, I., Torisawa, K. et De Saeger, S. (2010). Co-star : a co-training style algorithm for hyponymy relation acquisition from structured and unstructured text. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 842–850. Association for Computational Linguistics.

O'Leary P., D. (1996). Conjugate gradients and related kmp algorithms : the beginnings. *In* Adams, L. et Nazareth, J. L., éditeurs : *Linear and Nonlinear Conjugate Gradient-Related MethodsLinear and Nonlinear Conjugate Gradient-Related Methods*. SIAM, Philadelphia.

Omrane, N., Nazarenko, A. et Szulman, S. (2011). Le poids des entités nommées dans le filtrage des termes d'un domaine. *In International Conference on Terminology and Artificial Intelligence (TIA)*.

Paass, G. et Konya, I. (2012). Machine learning for document structure recognition. *In* Mehler, A., Kühnberger, K.-U., Lobin, H., Lüngen, H., Storrer, A. et Witt, A., éditeurs : *Modeling, Learning, and Processing of Text Technological Data Structures*, volume 370 de *Studies in Computational Intelligence*, chapitre Part V : Document Structure Learning, pages 221–247. Springer.

Palfray, T., Hebert, D., Nicolas, S., Tranouez, P. et Paquet, T. (2012). Logical segmentation for article extraction in digitized old newspapers. *In Proceedings of the 2012 ACM symposium on Document engineering*, pages 129–132. ACM.

Pantel, P. et Pennacchiotti, M. (2006). Espresso : Leveraging generic patterns for automatically harvesting semantic relations. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Pantel, P. et Pennacchiotti, M. (2008). Automatically harvesting and ontologizing semantic relations. *In* Buitelaar, P. et Cimiano, P., éditeurs : *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, volume 167 de *Frontiers in Artificial Intelligence and Applications*, pages 171–195. IOS Press, Amsterdam.

Pascual, E. (1991). *Représentation de l'architecture textuelle et génération de texte*. Thèse de doctorat, Université Paul Sabatier. Toulouse, France.

Pascual, E. et Péry-Woodley, M. (1995). La définition dans le texte. *Textes de type consigne–Perception, action, cognition*, 1:65–88.

Pascual, E. et Virbel, J. (1996). Semantic and layout properties of text punctuation. *In Proceedings of the Association for Computational Linguistics Workshop on Punctuation*, pages 41–48.

Pavlidis, T. et Zhou, J. (1991). Page segmentation by white streams. *In International Conference Document Analysis and Recognition (ICDAR)*, pages 945–953.

Peruzzo, K. (2014). Term extraction and management based on event templates : An empirical study on an eu corpus. *Terminology*, 20(2):151–170.

Phinney, T. W. (2004). Truetype, postscript type 1, & opentype : What's the difference. Rapport technique 2.36, Adobe.

Platt, J. *et al.* (1998). Sequential minimal optimization : A fast algorithm for training support vector machines.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *In Advances in large margin classifiers*. Citeseer.

Poibeau, T. (2005a). Parcours interprétatifs et terminologie. *In Actes du colloque Terminologie et Intelligence Artificielle*, page 12. Université de Rouen.

Poibeau, T. (2005b). Sur le statut référentiel des entités nommées. *In Conférence Traitement Automatique des Langues 2005*, pages 173–183. Association pour le Traitement Automatique des LAngues/LIMSI.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, 12(5):601–638.

Porhiel, S. (2007). Les structures énumératives à deux temps. *Revue romane*, 42(1): 103–135.

Power, R. (2000). Planning texts by constraint satisfaction. *In Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 642–648. Association for Computational Linguistics.

Power, R., Scott, D. et Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(2):211–260.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

Pustejovsky, J. (1995). *The Generative Lexicon : A Theory of Computational Lexical Semantics*. MIT Press.

Pustejovsky, J. et Stubbs, A. (2012). *Natural language annotation for machine learning*. O'Reilly.

Péry-Woodley, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Mémoire d'Habilitation à Diriger des Recherches (HDR), Université Toulouse le Mirail.

Péry-Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.-M., Le Draoulec, A. et Mathet, Y. (2009). Annodis : une approche outillée de l'annotation de structures discursives. *In Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*.

Péry-Woodley, M.-P. et Scott, D. (2006). Computational approaches to discourse and document processing. *TAL*, 47(2):7–19.

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. Elsevier.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ramakrishnan, C., Patnia, A., Hovy, E. H., Burns, G. A. *et al.* (2012). Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1):7.

Rangoni, Y. et Belaïd, A. (2006). Document logical structure analysis based on perceptive cycles. *In Document analysis systems VII*, pages 117–128. Springer.

Rastier, F. (1996). La sémantique des textes : concepts et applications. *Hermès*, 16:15–37.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *In Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA.

Ratté, S., Njomgue, W. et Ménard, P.-A. (2007). Highlighting document's structure. *International Journal of Computer Science & Engineering*, 1(2).

Ravi, S. et Paşca, M. (2008). Using structured text for large-scale attribute extraction. *In Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1183–1192. ACM.

Rebeyrolle, J. et Tanguy, L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, 25:153–174.

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H. et Wroe, C. (2004). Owl pizzas : Practical experience of teaching owl-dl : Common errors & common patterns. *In Engineering Knowledge in the Age of the Semantic Web*, pages 63–81. Springer.

Reichenberger, K., Kamps, T. et Golovchinsky, G. (1995). Towards a generative theory of diagram design. *In Information Visualization, 1995. Proceedings.*, pages 11–18. IEEE.

Reichenberger, K., Rondhuis, K. J., Kleinz, J. et Bateman, J. (1996). Effective presentation of information through page layout : a linguistically-based approach. *In Proceedings of ACM Workshop on Effective Abstractions in Multimedia, Layout and Interaction*, San Francisco. Association for Computing Machinery.

Richardson, S. D., Dolan, W. B. et Vanderwende, L. (1998). Mindnet : acquiring and structuring semantic information from text. *In Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1098–1102. Association for Computational Linguistics.

Riedel, S., Yao, L. et McCallum, A. (2010). Modeling relations and their mentions without labeled text. *In Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Rigau, G., Rodriguez, H. et Agirre, E. (1998). Building accurate semantic taxonomies from monolingual mrds. *In Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1103–1109. Association for Computational Linguistics.

Rizzo, G. et Troncy, R. (2012). Nerd : a framework for unifying named entity recognition and disambiguation extraction tools. *In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76. Association for Computational Linguistics.

Robinson, J. J. (1970). Dependency structures and transformational rules. *Language*, pages 259–285.

Role, F. et Rousse, G. (2006). Construction incrémentale d'une ontologie par analyse du texte et de la structure des documents. *Document numérique*, 9(1):77–91.

Rosario, B. et Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 430. Association for Computational Linguistics.

Rosenberg, A. et Binkowski, E. (2004). Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. *In Proceedings of HLT-NAACL 2004 : Short Papers*, pages 77–80. Association for Computational Linguistics.

Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rostek, L., Möhr, W., Fischer, D. H. *et al.* (1994). Weaving a web : the structure and creation of an object network representing an electronic reference work. *In Proceedings of Electronic Publishing*, volume 6, pages 495–506.

Sager, J. (1990). *A practical course in terminology processing*. John Benjamins Publishing, Amsterdam.

Sagot, B. et Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. *In TALN 2008-Traitement Automatique des Langues Naturelles*.

Sajous, F., Navarro, E. et Gaume, B. (2011). Enrichissement de lexiques sémantiques approvisionnés par les foules : le système wisigoth appliqué à wiktionary. *Traitement Automatique des Langues*, 52(1):11–35.

Sammut, C. et Webb, G. (2010). *Encyclopedia of machine learning*. Springer.

Sarawagi, S. et Cohen, W. W. (2004). Semi-markov conditional random fields for information extraction. *In NIPS*, volume 17, pages 1185–1192.

Schnedecker, C. (2002). Adverbes ordinaux et introducteurs de cadre : aspects linguistiques et cognitifs. *Lingvisticae Investigationes*, 24(2):257–287.

Schrod, J. (1991). The components of tex. Rapport technique, Detig Schrod TEXsys.

Schürmann, J., Bartneck, N., Bayer, T., Franke, J., Mandler, E. et Oberländer, M. (1992). Document analysis-from pixels to contents. *Proceedings of the IEEE*, 80(7):1101–1119.

Scott, D. et de Souza, C. S. (1990). Getting the message across in rst-based text generation. *Current research in natural language generation*, 4:47–73.

Scott, W. A. (1955). Reliability of content analysis : The case of nominal scale coding. *Public opinion quarterly*.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, 5(01):1–23.

Sébillot, P. (2002). *Apprentissage sur corpus de relations lexicales sémantiques-La linguistique et l'apprentissage au service d'applications du traitement automatique des langues*. Thèse de doctorat, Université Rennes 1.

Séguéla, P. (1999). Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. *Terminologies nouvelles*, 19:52–60.

Séguéla, P. (2001). Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. *Thèse de doctorat, Université de Toulouse*.

Sha, F. et Pereira, F. (2003). Shallow parsing with conditional random fields. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics.

Shafait, F., Keysers, D. et Breuel, T. M. (2008). Performance evaluation and benchmarking of six-page segmentation algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):941–954.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal,*, 27(1):379–423.

Shinzato, K. et Torisawa, K. (2004a). Acquiring hyponymy relations from web documents. *In HLT-NAACL*, pages 73–80.

Shinzato, K. et Torisawa, K. (2004b). Extracting hyponyms of prespecified hypernyms from itemizations and headings in web documents. *In Proceedings of the 20th international conference on Computational Linguistics*, page 938. Association for Computational Linguistics.

Snow, R., Jurafsky, D. et Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *In Advances in Neural Information Processing Systems*, volume 17.

Southall, R. (1989). Interfaces between the designer and the document. *In* André, J., Furuta, R. et Quint, V., éditeurs : *Structured Documents*, pages 161–180. Cambridge Series on Electronic Publishing.

Stubbs, A. (2011). Mae and mai : Lightweight annotation and adjudication tools. *In Proceedings of the 5th Linguistic Annotation Workshop, Association of Computational Linguistics*, Portland.

Suchanek, F. M., Kasneci, G. et Weikum, G. (2007). Yago : a core of semantic knowledge. *In Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Sumida, A. et Torisawa, K. (2008). Hacking wikipedia for hyponymy relation acquisition. *In IJCNLP*, volume 8, pages 883–888.

Sumida, A., Yoshinaga, N. et Torisawa, K. (2008). Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia. *In Proceedings of the 6th International Conference on Language Resources and Evaluation.*

Sutton, C. et McCallum, A. (2006). *An introduction to conditional random fields for relational learning*, volume 2. Introduction to statistical relational learning. MIT Press.

Tadros, A. (1985). *Prediction in text.* Numéro 10. English Language Research.

Tang, Y. Y., Lee, S.-W. et Suen, C. Y. (1996). Automatic document processing : a survey. *Pattern recognition*, 29(12):1931–1952.

Tang, Y. Y., Ma, H., Mao, X., Liu, D. et Suen, C. Y. (1995). A new approach to document analysis based on modified fractal signature. *In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 2, pages 567–570. IEEE.

Tateisi, Y., Ohta, T., Collier, N., Nobata, C. et Tsujii, J.-i. (2000). Building an annotated corpus in the molecular-biology domain. *In Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, pages 28–36. Association for Computational Linguistics.

Tesnière, L. (1959). *Eléments de syntaxe structurale.* Librairie C. Klincksieck.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D. et Manning, C. (2005). A conditional random field word segmenter for sighan bakeoff 2005. *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 171. Jeju Island, Korea.

Tsujimoto, S. et Asada, H. (1992). Major components of a complete text reading system. *Proceedings of the IEEE*, 80(7):1133–1149.

Turco, G. et Coltier, D. (1988). Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire. *Pratiques*, 57:57–79.

Turner, J. (1996). *The Dictionary of Art*. Macmillan Publisher, New York.

Urieli, A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer, New York.

Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Vapnik, V. et Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, 24:774–780.

Venetis, P., Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G. et Wu, C. (2011). Recovering semantics of tables on the web. *Proceedings of the VLDB Endowment*, 4(9):528–538.

Vergez-Couret, M., Bras, M., Prevot, L., Vieu, L., Attalah, C. *et al.* (2011). The discourse contribution of enumerative structures involving 'pour deux raisons'. *In Procceedings of Constraints in Discourse*.

Vergez-Couret, M., Prévot, L. et Bras, M. (2008). Interleaved discourse, the case of two-step enumerative structures. *In Proceedings of Contraints In Discourse III*, pages 85–94, Potsdam.

Véronis, J. et Ide, N. (1991). An assessment of semantic information automatically extracted from machine readable dictionaries. *In Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, pages 227–232. Association for Computational Linguistics.

Viola, P. et Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511. IEEE.

Virbel, J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de grammaire*, (10):5–72.

Virbel, J. (1989). The contribution of linguistic knowledge to the interpretation of text structures. *In* André, J., Furuta, R. et Quint, V., éditeurs : *Structured Documents*, pages 161–180. Cambridge Series on Electronic Publishing.

Virbel, J. (1999). Structures textuelles, planches fascicule 1 : Enumérations, version 1,. Rapport technique, IRIT.

Virbel, J., Luc, C., Schmid, S., Carrio, L., Dominguez, C., Péry-Woodley, M.-P., Jacquemin, C., Mojahid, M., Baccino, T. et Garciadebanc, C. (2005). Approche cognitive de la spatialisation du langage. de la modélisation de structures spatio-linguistiques des textes à l'expérimentation psycholinguistique : le cas d'un objet textuel, l'énumération. *In* Thinus-Blanc, C. et Bullier, J., éditeurs : *Agir dans l'Espace*, chapitre 12, pages 233–254. Paris : Editions de la MSH.

Völker, J., Vrandečić, D., Sure, Y. et Hotho, A. (2007). Learning disjointness. *In The Semantic Web : Research and Applications*, pages 175–189. Springer.

Vossen, P. (1998). Eurowordnet : building a multilingual database with wordnets for european languages. *The ELRA Newsletter*, 3(1):7–10.

Warnock, J. (1991). The camelot project. *Adobe*.

Wells, R. S. (1947). Immediate constituents. *Language*, 23(2):81–117.

Widlöcher, A. et Mathet, Y. (2009). La plate-forme glozz : environnement d'annotation et d'exploration de corpus. *In Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*.

Wierzbicka, A. (1972). Semantic primitives.

Wilks, Y. (1993). Providing machine tractable dictionary tools. *In* Pustejovksy, J., éditeur : *Semantics and the Lexicon*, pages 341–401. Springer.

Williams, C. K. (1998). Prediction with gaussian processes : From linear regression to linear prediction and beyond. *In Learning in graphical models*, pages 599–621. Springer.

Winograd, T. (1978). On primitives, prototypes, and other semantic anomalies. *In Proceedings of the 1978 workshop on Theoretical issues in natural language processing*, pages 25–32. Association for Computational Linguistics.

Wolpert, D. H. et Macready, W. G. (1997). No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82.

Wong, K. Y., Casey, R. G. et Wahl, F. M. (1982). Document analysis system. *IBM journal of research and development*, 26(6):647–656.

Woods, W. A. (1975). What's in a link : Foundations for semantic networks. *Representation and understanding : Studies in cognitive science*, pages 35–82.

Wüster, E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. *Textes choisis de terminologie, GIRSTERM, Université de Laval, Québec*, pages 55–108.

Xu, W., Hoffmann, R., Zhao, L. et Grishman, R. (2013). Filling knowledge base gaps for distant supervision of relation extraction. *In ACL (2)*, pages 665–670.

Yamada, I., Oh, J.-H., Hashimoto, C., Torisawa, K., Kazama, J., De Saeger, S. et Kawada, T. (2011). Extending wordnet with hypernyms and siblings acquired from wikipedia. *In IJCNLP*, pages 874–882.

Yamada, I., Torisawa, K., Kazama, J., Kuroda, K., Murata, M., De Saeger, S., Bond, F. et Sumida, A. (2009). Hypernym discovery based on distributional similarity and hierarchical structures. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2-Volume 2*, pages 929–937. Association for Computational Linguistics.

Yao, L., Riedel, S. et McCallum, A. (2012). Unsupervised relation discovery with sense disambiguation. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pages 712–720. Association for Computational Linguistics.

Yoshinaga, N. et Torisawa, K. (2007). Open-domain attribute-value acquisition from semi-structured texts. *In Proceedings of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge : The Lexicon/Ontology Interface (OntoLex-2007)*, pages 55–66.

Zargayouna, H. (2004). Contexte et sémantique pour une indexation de documents semi-structurés. *CORIA*, 4:161–177.

Zelenko, D., Aone, C. et Richardella, A. (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106.

Zesch, T., Müller, C. et Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. *In Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco. electronic proceedings.

Zhang, C., Xu, W., Gao, S. et Guo, J. (2014). A bottom-up kernel of pattern learning for relation extraction. *In Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, pages 609–613. IEEE.

Zhao, S. et Grishman, R. (2005). Extracting relations with integrated information using kernel methods. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 419–426. Association for Computational Linguistics.

Zhou, Z.-H. (2012). *Ensemble methods : foundations and algorithms*. CRC Press.